SOME INVESTIGATIONS ON SINGLE AND MULTI-CHANNEL BLIND SOURCE SEPARATION USING ARTIFICIAL INTELLIGENCE TECHNIQUES

Submitted in partial fulfilment of the requirements of the degree of

DOCTOR OF PHILOSOPHY

by

YANNAM VASANTHA KOTESWARARAO

(Roll No.717027)

Under the guidance of

Prof. C. B. Rama Rao

Professor, Dept. of

ECE

DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING

NATIONAL INSTITUTE OF TECHNOLOGY WARANGAL

TELANGANA, INDIA-506 004

2023

APPROVAL SHEET

This Dissertation Work entitled **"Some Investigations on Single and Multi-channel Blind Source Separation using Artificial Intelligence Techniques"** by **Yannam Vasantha Koteswararao,Roll No. 717027** is approved for the degree of Doctor of Philosophy.

Examiners

_____

_____

_____

Supervisor

_____

Chairman

_____

Date:_____ Place: _____

DECLARATION

This is to certify that the work presented in the thesis entitled **"Some Investigations on Single and Multi-channel Blind Source Separation using Artificial Intelligence Techniques"** is a bonafide work done by me under the supervision of **Prof. C.B. RamaRao**, **Professor,** Department of Electronics and Communication Engineering, National Institute of Technology, Warangal and was not submitted elsewhere for the award of any degree.

I declare that this written submission represents my ideas in my own words and where others 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 / data / fact / source in my submission. I understand that any violation of the above will be a 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.

Yannam Vasantha Koteswararao

(Roll No. 717027)

Place: Warangal

Date:

DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING
NATIONAL INSTITUTE OF TECHNOLOGY, WARANGAL
TELANGANA, INDIA-506 004

CERTIFICATE

This is to certify that the thesis entitled "**Some Investigations on Single and Multi-channel Blind Source Separation using Artificial Intelligence Techniques**" is being submitted by **Mr. Yannam Vasantha Koteswararao (Roll No. 717027)**, in partial fulfilment 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 him under my supervision and guidance and has not been submitted elsewhere for any degree.

Signature of Supervisor

Prof. C.B. Rama Rao
Professor
Department of ECE,
National Institute Technology,
Warangal, Telangana,
India- 506004.

ACKNOWLEDGEMENTS

It gives me immense pleasure to convey my deep sense of gratitude and sincere thanks to my supervisor **Prof. C.B. Rama Rao**, Professor, Department of ECE, NIT, Warangal, for their perpetual encouragement, guidance, and supervision. Their steady influence throughout my Ph.D. career has oriented me in a proper direction and supported me with promptness and care. They not only gave me the required knowledge to pursue my research work, but also gave me the required moral support during my hard times. I truly appreciate their logical and thought provoking advises both technically and morally which I will follow for the rest of my life.

I am also grateful to **Prof. Patri Sreehari Rao**, Head, Department of Electronics and Communication Engineering, for his invaluable assistance and suggestions that he shared during my research tenure

Also, I take this privilege to thank all my Doctoral Scrutiny Committee members, **Prof. V. Venkata Mani**, Professor, Head Department of ECE, **Prof. S. Anuradha**, Professor, Department of ECE, **Prof. Y. N. Reddy**, Professor, Mathematics Department, for their detailed review, constructive suggestions, and excellent advice during the progress of this research work.

I am also thankful to the former Heads of the ECE department **Prof. T. Kishore Kumar, Prof. N. Bheema Rao and Prof. L. Anjaneyulu** for their continuous support and motivation. I thank all the faculty and non-teaching staff of ECE **Dept. at NIT Warangal** who helped me during the tenure of my research work.

I am also grateful to all my colleagues, scholars, friends, and well-wishers who helped me to write my thesis with their support. Also, especially I would like to thank Dr. P. Muralidhar, Dr. V. Rama, Dr. R. K. Niranjan, Mr. S. Siva Prasad, Mr. K. Rambabu and Mr. P. Hari for their help and support during my Ph.D.

I would like to acknowledge my biggest debt to my family members for their continuous support.

Yannam Vasantha Koteswararao

ABSTRACT

Speech signal processing has been one of the domains of research after the past one decade in signal processing. Research has taken new strides particularly during the past decade (Five to Ten years). In an environment where multiple speech signals are generated from different know or unknown sources, they may be mixed-up with various background noise sources, reverberations and interference, Accordingly the terms like target sources and blind sources may be used while dealing with them in such environment.

Blind source separation (BSS) is one of the challenging problems in speech signal processing. When a single microphone is used to sense the sources, it is referred as single channel and when two or microphones are sensing it is called multichannel, but ultimately it may be of interest to extract/ enhance a single desirable speech signal reserving its quality related matrix and thus eliminating all the rest. The desirable speech signal may be termed as target sources and all the rest blind sources. The target sources may be masked by other interfering blind sources as well as corrupted by various background noise sources and reverberations hence blind sources separation is required for enhancement extraction of target sources.

This work has been focused on development of optimized matrix factorization integrated with deep Learning methods for BSS. When the sources are very much limited, the problem of BSS will be simpler. However, when the sample size of sources is reasonably large, the problem will become complex. The methods proposed in this work deal with complex situations. In the proposed research problem one or more than one speech signal mixed with different types of WSS noise sources has been considered.

This proposed research work consists of four contributions for single and multi-channel source separation. They are:

i) Time-Frequency Non-Negative Matrix Factorization (TFNMF) and Sigmoid Base Normalization Deep Neural Networks for Single Channel Source Separation. Experiments show that our proposed method achieves the highest gains in PESQ, STIO, SIR and SDR whose numerical values are 3.58, 0.7, 42 and 7.5 at -9 dB. These obtained results are compared with those of existing works.

ii) Integral fox ride optimization (IFRO) algorithm and retrieval-based deep neural network (RDNN) for Single Channel Source Separation. Experiments show that our proposed method achieves the highest gains in SDR, SIR, SAR STIO, and PESQ whose numerical values are 10.9, 15.3, 10.8, 0.08, and 0.58, respectively. The Joint-DNN-

SNMF obtains 9.6, 13.4, 10.4, 0.07, and 0.50, comparable to the Joint-DNN-SNMF. These obtained results are compared with those of existing works.

iii) Grasshopper Optimization-based Matrix Factorization (GOMF) and Enthalpy based DNN (EDNN) for multichannel source Separation. Experimental results show that our proposed approach accomplishes the most extreme SNR outcome of − 6dB of 24.0523. Comparable to the DNN-JAT, which achieves 18.50032. The RNN and NMF-DNN had the worst SNR 13.45434 and 12.29991. These obtained results are compared with those of existing works.

iv) krill herd-based matrix factorization (KHMF) and score-based convolutional neural network (SCNN) for multichannel Source Separation. Experimental results show that our proposed approach accomplishes the most extreme SDR dif outcome of − 5 dB of 8.1. Comparable to the CTF-MINT, which achieves 8.05. The CTF-MPDR and CTF-BP had the SDR dif worst 7.71 and 7.4. The Unproc had the very worst SDR dif 5.71. These obtained results are compared with those of existing works.

All the proposed source separation models are evaluated for the mixed sources. The investigation has been carried out experiments are carried out with various data sets. The standard source evaluation objective parameters, such as signal to distortion ratio (SDR), signal to interference ratio (SIR), perceptual evaluation of speech quality (PESQ), short time objective intelligibility (STOi) and signal to artefacts ratio (SAR), are used for ensuring the quality of enhancement.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

Abbreviations	Description
MME	Modulation Magnitude Estimator
MMSE	Minimum Mean Squared Error
BSS	Blind Source Separation
STFA	Short-Time Fourier Analysis
DNN	Deep Neural Network
ERSS	Efficient Optimal Reconstruction-Based Speech Separation
IFRO	Integral Fox Ride Optimization
RDNN	Retrieval-Based Deep Neural Network
SS	Speech Separation
GOMF	Grasshopper Optimization-Based Matrix Factorization
KHMF	Krill Herd-Based Matrix Factorization
GMM	Gaussian Mixture Model
SCNN	Score-Based Convolutional Neural Network
DOA	Enthalpy-Based Direction Of Arrival
ASR	Automatic Speech Recognition
SCSS	Single-Channel Speech Separation
MCSS	Multi-Channel Speech Separation
MISO	Multiple Input One Output System
CASA	Computational Auditory Scene Analysis
DESA	Discrete Energy Separation Algorithm
NMF	Non-Negative Matrix Factorization
dFDLR	Discriminant Linear Regression
DANet	Deep Attractor Network
HVQHCA	Hybrid Vector Quantization-Based Heuristic Clustering Algorithm
SDR	Signal To Distortion Ratio
VGG	Visual Geometry Group
ResNets	Residual Neural Networks
ICA	Independent Component Analysis
ICDs	Inter-Channel Convolution Differences

TAC	Transform-Average-Concatenate
PIT	Permutation Invariant Training
SCBSS	Single-Channel Blind Source Separation
TFNMF	Time Frequency Non-Negative Matrix Factorization For Source Separation

Chapter 1
Introduction

1.1 General

In the actual world, voice signals are recorded using a single microphone or a number of microphones and then transferred to computers for additional processing. Multiple microphones are obviously desired during the collection process, if the circumstances allow. In this situation, spatial cues can be kept and utilised as additional tools for deciphering mixed speech. If the target speaker is not predetermined, microphone arrays may not be advantageous in cocktail party settings with many sound sources. Even worse, there isn't always access to a setting that allows for numerous mics. When it comes to automatic speech recognition for radio broadcasts, utilising one microphone is frequently the only option. Since there is no location information available, voice activity in this scenario is sent and recorded through radio channels. The news anchor's voice is frequently distorted by background speakers. Teleconferences are another real-time use of speech recognition.

If simultaneous speech is recorded and delivered to a speech recognizer, the accuracy is poor. Any modern recognizer finds the task extremely challenging when there are multiple interfering speakers present. The output of the recognition system is typically subjected to additional processing by these systems, including text-to-speech, dialogue systems, question-answering, news summarising, and categorization. For all of these applications to attain a reasonable level of speech recognition accuracy, a good single-channel speech system is necessary because a low level of speech recognition accuracy could result in a substantial accumulation of errors (Daniel et al 2004, 2007) [1]. The issue of single and multi-channel speech recognition in interfering noise must be solved for these reasons, and it is crucial.

The classic issue in auditory scene analysis is sound source separation. The difficult issue of extracting individual voice streams from a mixed signal of several speakers using single and multi-channel speech separation, in particular, has applications in reliable automatic speech recognition, speech augmentation, and other areas [2-12].

The voice signal has been improved utilising the minimum mean squared error (MMSE) - Short Time Spectral Magnitude approach by Wang et al. (2014) [13]. This method is used to determine starting parameters and the modulation spectrum for noise distortion. Comparatively speaking to the many other improvement methods, this improves the subjective quality of noisy speech across multiple acoustic domains. The Modulation Magnitude Estimator (MME)

parameters are used in several tests to optimise the enhanced speech quality while minimising noise distortion.

1.2 Blind Source Separation

In order to identify each signal element within the blending made up of numerous sensing units, the Blind Source Separation (BSS) method is generally used. It is referred to as blind because no information other than the combinations is employed. In a hall, for instance, a group of people is speaking, and microphones are being utilised to record the signals [14]. When one or more people are chatting at the same time, the electro acoustical sensor of each speaker records a variety of blending as the voice signals are logged for each person separately. Currently, BSS must finish the work of disentangling such blending from its original supply signals, which are the voice inputs of each individual speaking. It is challenging generally because of some complication issues [15].

1.3 Speech Mixture Generation

The process of human communication known as speaking is generally used. The human auditory system can distinguish between the target sound and background noise interference. However, many disturbances like train noise, fan noise, crowd noise, etc., interfere with this communication. The separation of monaural speech remains one of the most difficult issues in speech processing, and numerous solutions have been put forth to address this issue. In order to improve speech that has been corrupted by additive non-speech noise, speech enhancement techniques make use of the statistical characteristics of the signal [16-22]. For voice improvement, noise reduction, and improving the quality and understandability of speech, various research activities have been offered.

Two distinct speech signals produced by two different speakers of the same gender and a different gender are used as the system's input. The two speech signals are subjected to feature extraction, which includes pitch values, phase, angle, and fundamental frequency (F0). To create a combination of speech signals, both signals are mixed. For further processing, this mixed speech signal is employed. X[n] represents the speech mixture. Given that a[n] and b[n] are two distinct speech signals, x[n] is given as

$$X[n]= aa[n] + b[n] \quad (1.1)$$

It should be noted that the two voice signals are added together without being scaled in any way.

1.3.1 Speech Separation

In order to solve the speech separation problem, a signal that contains a target source, an interference source, reverberations, and noise when it reaches the receiver must be separated out to achieve the desired speech source. Processing ought to keep the intended voice source and throw away the rest of the signal [23]. Given that there are numerous voice sources in the sound area, the target source could be any one of them or all of them. This stands out from the vast majority of single target source improvement problems, also referred to as the traditional speech denoising problems.

1.3.2 Signal channel Separation Process

Short-time Fourier analysis (STFA), where m' is the frame index, n' is the time sample index inside a frame, and 'k' is the index of frequency bins, is used to first decompose the combined speech X[n] into a two-dimensional time-frequency representation. Both of the parallel signal separation methods supported by the system—one based only on the fundamental frequency and the other on correlations of modulation frequency—use the same speech recognition and peripheral signal processing software.

Underdetermined blind source separation occurs when n < m. It is known as single-channel blind source separation under uncertain conditions when n = 1. The instantaneous mixing model for single-channel underdetermined blind source separation is as follows:

$$y(t) = \sum_{i=1}^{N} a_i e_i\,(t) \qquad (1.2)$$

1.3.3 Multi-channel speech separation

When multiple speech signals are collected using a single microphone or when multiple speech signals are delivered through the same communication channel, the process is known as multi-channel speech separation. Many speeches processing applications, including automatic speech recognition, speaker recognition, audio retrieval, and hearing aids, can be considerably aided by such tasks [24-26]. In the forensics division, the speech mixture that was recorded along with the video capture can be divided into different speech signals and examined.

The wave shape of the observed signal y(t) and the independence between the signal sources are employed in the blind source separation to get the estimated signal e* (t) as near to the signal source e(t) as feasible. Linear instantaneous mixing model is the mathematical representation of blind source separation [27-35]:

$$y(t) = Me(t) + nm \qquad (1.3)$$

M stands for the mixing matrix in the equation, whereas m and n stand for the number of source signals and receiving antenna components, respectively.

1.3.3.1 Application using BSS

Blind source separation (BSS) is a signal processing technique used to separate independent signals from a mixture of signals. BSS has a wide range of applications in various fields, including:

- Speech and audio processing: BSS can be used to separate different sources of speech or music from a mixed audio signal. This is useful in applications such as noise reduction, speaker separation, and audio signal enhancement.
- Image processing: BSS can be used to separate different sources of images from a mixed image signal. This is useful in applications such as object detection and image segmentation.
- Biomedical signal processing: BSS can be used to separate different sources of physiological signals, such as electrocardiogram (ECG) and electroencephalogram (EEG) signals. This is useful in applications such as diagnosing heart diseases and brain disorders.
- Radar and sonar signal processing: BSS can be used to separate different sources of radar or sonar signals from a mixed signal. This is useful in applications such as target tracking and detection.
- Financial data analysis: BSS can be used to separate different sources of financial data, such as stock prices and economic indicators. This is useful in applications such as portfolio management and risk analysis.

Overall, BSS is a powerful technique that has a wide range of applications in various fields where it is necessary to separate independent signals from a mixture of signals.

1.4 Research Motivation

The input speech signal is frequently distorted by the ambient acoustic noise in many speeches processing applications, including speaker identification, speech enhancement, and speech recognition. This ultimately lowers the perceived quality and understandability of the speech, which lowers the overall effectiveness of the speech processing system. In order to improve voice quality and understandability for future processing, a speech separation algorithm serves as a crucial front-end component [36]. It will improve the overall performance of the speech processing algorithm if the desired speech signal is extracted from the acoustic sounds before processing. Due to the ease of installing a microphone, there may only be one

acquisition channel available in some real-world circumstances. However, the main drawback of single channel approaches is the lack of a reference signal to compare interference signals against. Because of this, it is difficult to quantify the power spectral density of the interfering speech using the multi-channel speech signals that are currently available. The reduction of artefacts in the processed speech is crucial, particularly if the recovered speech is intended to be used in machine-based applications like speaker identification and automatic speech recognition.

1.5 Problem Statement

Over the past decade, there has been substantial research in the area of speech separation from various noise sources and interference. The desired voice signal may be corrupted by noise in an additive or multiplicative way whose spectrum is constant. The desired single voice signal could also be joined by some additional interfering sources, such as multiple speech signals. All of them are commonly referred to as "blind sources." The prerequisite is the suppression or cancellation of noise, as well as the separation of all sources of unwanted interference—aside from the intended voice signal—from one other. This improvement is accomplished by separating blind sources. In the area of the problem indicated above, several scholars have tried a few different approaches. The works stated in chapter 2 that have been modified or used unique techniques to increase the quality of the augmented speech signal. Modifications to techniques like TF, NMF, and deep learning testing models are used in this research. Both single-channel and multi-channel speech separation techniques are taken into consideration in this research.

1.6 Motivation

The speech separation issue is the focus of this thesis in both single- and multi-channel scenarios, both supervised and unsupervised. According to this fundamental premise, the original signal will be handled as a mixture in which both the desired and undesirable speech signal components are present. The suggested approaches will be used to improve the recorded mixed signal, maintaining the required components and deleting the undesirable ones. The target speech source is the desired component, whereas background noise, reverberation, and interfering speech sources make up the undesirable component.

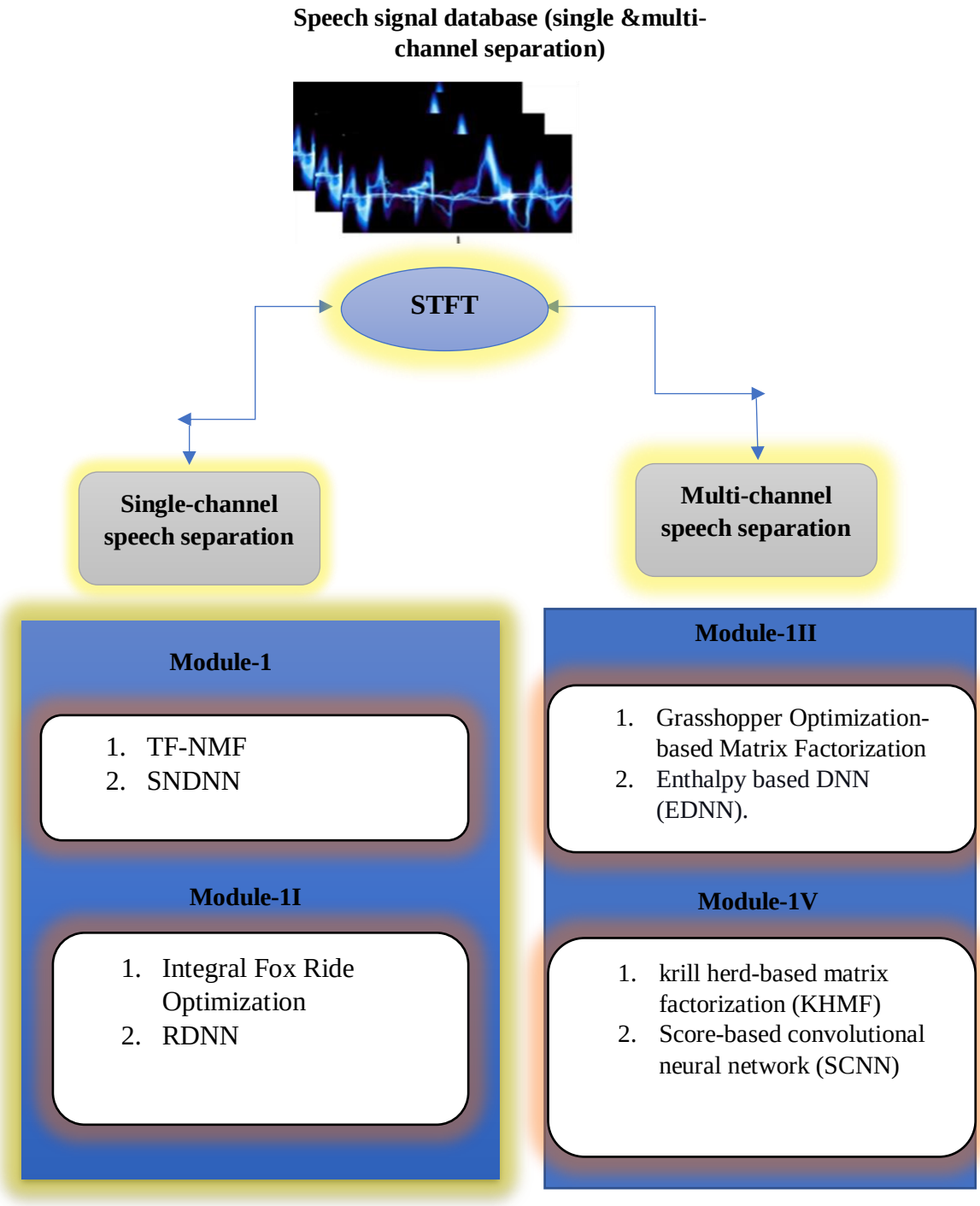
The second chapter contains a detailed description of the contributions made by numerous researchers. For source separation in single channel and multi-channel environments, a variety of strategies have been put forth by various scholars. Despite the fact that NMF-based techniques produce superior outcomes, rank estimation remains a significant challenge for modern NMF systems. The use of approximation signals to increase the training data set and over-smoothing are noted to be two long-standing problems with single-channel speech separation. Additionally, when many signals are separated from their superposition recorded at different sensors in multi-channel environments, problems such as insufficient separation and voice distortion mitigation have emerged. Finding the eigenvalues of a noise signal in a multichannel situation is a similar task.

A two-fold mistake occurs when each activation is carried out separately, making the deviation more susceptible to errors in Deep Neural Network approximation. As a result, the problem of the spectral overlay at the beginning of a dialogue or commotion is lessened and discriminative grounds are created. On the other hand, there are still several issues, including a lack of robustness and inadequate separation accuracy. In order to increase performance parameters like SAR, SDR, SIR, SNR, and PESQ, we must thus present novel methodologies based on the deep learning idea in order to address the aforementioned issue.

1.7 Proposed system flow

The thesis' main goal is to offer some input into the design and implementation of a reliable and improved single channel and multi-channel speech separation system in a clean and noisy environment with reduced time complexity, which can be used in practical applications for hard of hearing people and in forensic departments to identify the speech recorded in public places. The suggested separation system's framework is depicted schematically in Figure 1.1.

Figure :1.1 Propose framework of Speech Separation System



Both single channel and multi-channel speech signal separation are goals of the study activity. In the proposed work consist of four modules, in first two module focused only on single channel speech separation using neural network and also last two module is implemented in multi-channel speech signal with hybrid neural networks. which is,

- **Module 1:** To categorise single-channel source segmentation using sigmoid-based normalisation in addition to deep neural networks with time-frequency non-negative matrix factorization is one of the study's primary goals.
- **Module 2:** To enhance the training data set's quality by employing a hybrid deep learning approach in a single channel environment to extract low-level texture information from each spoken signal.
- **Module 3:** Using Optimization with Matrix Factorization and DNN to categorise a multichannel voice input.
- **Module 4:** To classify an audio signal using a novel hybrid approach that uses a convolutional neural network (SCNN).

Module 1:

The two steps that make up the suggested system are listed below. The training phase comes first, and the testing phase follows. The testing stage employs a single-channel multi-talker input signal, whereas the training stage uses a single-channel clean input signal. This distinction between the two testing and training phases allows for more accurate comparisons. The input signals from these testing and training phases are sent to the short-term fourier transform (STFT). When extracting features from spectrograms created by STFT, which transforms input clean signal into spectrograms, TFNMF is the approach used. Utilizing the SNDNN classification algorithm after feature extraction, the classified features are then converted to softmax. Then, ISTFT utilises softmax to appropriately separate speech signals.

Module 2:

Conventional single-channel speech separation has two long-standing issues. The first issue, over-smoothing, is addressed, and estimated signals are used to expand the training data set. Second, DNN generates prior knowledge to address the problem of incomplete separation and mitigate speech distortion [37]. To overcome all current issues, we suggest employing an efficient optimal reconstruction-based speech separation (ERSS) to overcome those problems using a hybrid deep learning technique. First, we propose an integral fox ride optimization (IFRO) algorithm for spectral structure reconstruction with the help of multiple spectrum features: time dynamic information, binaural and mono features. Second, we introduce a hybrid

retrieval-based deep neural network (RDNN) to reconstruct the spectrograms size of speech and noise directly. The input signals are sent to Short Term Fourier Transform (STFT). STFT converts a clean input signal into spectrograms then uses a feature extraction technique called IFRO to extract features from spectrograms. After extracting the features, using the RDNN classification algorithm, the classified features are converted to softmax. ISTFT then applies to softmax and correctly separates speech signals. Experiments show that our proposed method achieves the highest gains in SDR, SIR, SAR STIO, and PESQ outcomes of 10.9, 15.3, 10.8, 0.08, and 0.58, respectively. The Joint-DNN-SNMF obtains 9.6, 13.4, 10.4, 0.07, and 0.50, comparable to the Joint-DNN-SNMF. The proposed result is compared to a different method and some previous work. In comparison to previous research, our proposed methodology yields better results.

Module 3:

In real environments, room reverberation and associated sounds frequently degrade speech transmission. This study focuses on decoupling objective speech signals from multichannel input sources under reverberant circumstances. This work presents an effective method for multichannel speech signal separation utilising a new hybrid technique that combines Grasshopper Optimization-based Matrix Factorization (GOMF) with Enthalpy-based DNN in order to address all the current shortcomings (EDNN).

This research proposes a narrative classification framework that includes the phases of STFT, GOMF-based rank estimation, identifying signal Eigenvalues, noise reduction, feature extraction, and classification in order to forecast and remove the undesirable noise from the multichannel input signal. The multichannel mix waveforms are first planned using STFT to create complex spectrograms. The evident speech signals and noise are then estimated using GOMF. Important features are extracted after the estimation. Spatial, spectral, and directional features serve as the foundation for feature extraction. A deep neural network based on enthalpy is used to recreate the spectrogram in order to achieve improved classification results (EDNN).

Finally, using inverse STFT, transform the generated speech spectrogram back into the retrieved output signal. According to experimental findings, our suggested method achieves the highest extreme SNR result, a -6dB of 24.0523. comparable to the 18.50032 achieved by the DNN-JAT. The worst SNR values were 13.45434 and 12.29991 for the RNN and NMF-DNN.

Module 4

Multi-channel speech separation (SS) is the process of isolating a multi-channel speaker's voice from the simultaneous speaker's overlapping sounds. Visual modalities have so far demonstrated considerable promise for multi-channel speech separation. It is addressed how to separate multiple signals from their superposition when they are recorded at several sensors. The use of a novel hybrid method combining enthalpy-based direction of arrival (DOA) and krill herd-based matrix factorization (KHMF) to segment multi-channel speech signals, as well as Convolutional neural network (SCNN) estimation, are some of the solutions this article suggests to address any current drawbacks. First, determine the input signal's short-term Fourier transform (STFT). The tracking branch then starts to determine the signal's enthalpy after signal analysis. The spatial energy based on DOA in each time frame is known as enthalpy. The spatial energy histogram is converted into DOA measurements by the Gaussian Mixture Model (GMM), which also calculates the enthalpy function at each time frame. The output of the signal tracker is used to determine an enthalpy-based spatial covariance matrix model with DOA parameters [38]. Utilize multi-channel KHMF to calculate the source's spectral model and spatial behaviour over time from the tracking direction. Effective qualities like directivity and spatial features are then extracted based on the target speaker's spatial direction. Utilize the relation masking function of the score-based convolutional neural network (SCNN). The extracted output signal is converted from the generated speech spectrogram using the STFT (iSTFT) procedure. According to experimental findings, our suggested method achieves the most extreme SDR diff result, which is -5dB of 8.1. comparable to the CTF-8.05 MINT's score The SDR diff for the CTF-MPDR and CTF-BP were 7.71 and 7.4, respectively. SDR diff 5.71 for the Unproc was the worst possible.

NMF is used to understand the important spectrum of speech and sound, whereas DNN is used to assess the essential spectrum's function. The NMF hypothesis and functional assessment are combined with DNN to comprehensively reproduce clear sound and sound within the compound. The combined strains of DNN and NMF are improving the performance of the voice department. We suggest a different optimization range with interval control to suppress excessive noise. This reduces the residue of isolated speech and noise and dramatically improves GSIR performance. Models can stop high interactions and outperform comparative models with very low-cost hand tools and defects. Production models can use spectral structures based on speech and sound, while in-depth study models study complex linear graphs of distinguishing objectives through silent and supervised learning.

The latest approach since optimizes training formal speech segmentation, in which different modes of speech, speaker, and background sound are studied from training data.

Several supervised separation systems have been proposed. The in-depth learning methods used for supervised speech separation increased the rate of progress and increased the separation efficiency. Also, reliable assessment of time-frequency masks from the conversation is challenging, especially when there is room echo in the mix.

We propose an efficient optimal reconstruction-based speech separation (ERSS) to overcome those problems using a hybrid deep learning technique.

- First, to compute the signal's STFT and computing the enthalpy of the signal.
- Second, propose an integral fox ride optimization (IFRO) algorithm for spectral structure reconstruction with the help of multiple spectra features: time dynamic information, binaural and mono features.
- Third, introduce the Deep Neural Network (RDNN) based on a hybrid search to directly reproduce the speech and voice level spectrogram. RDNN can instantly improve the partitioning range and minimize accumulated errors.
- • The GMM, which calculates the enthalpy function for each time frame. The monitored address is estimated using multi-channel KHMF, and the enthalpy DOA is utilised to parameterize the SCM model.
- • Following that, the spectrogram speech separation will be muted based on the SCNN score.
- Finally, we implement the proposed design in the MATLAB tool, and the performance of the proposed system is compared with the existing state-of-art techniques.

1.8 Contribution

When several speakers are speaking at almost the same time, speech separation is used to highlight each speaker's mixed-language discourse. It is helpful in speech-related systems because it can denoise, extract, and improve speech signals. In recent years, a variety of techniques to distinguish human voices from background noise and other noises have been put forth. The use of approximation signals to increase the training data set and over-smoothing are two long-standing problems with traditional single-channel speech separation. Inadequate separation and voice distortion mitigation have also become a problem. Single-channel source separation with Time-Frequency non-negative matrix factorization, sigmoid-based normalisation deep neural networks, and an effective optimal reconstruction-based speech separation (ERSS) method using a hybrid deep learning technique have all been developed to address all identified challenges.

The technique of separating the voice of a multi-channel speaker from the overlapping audio of a simultaneous speaker is known as multi-channel speech separation. Difficulties

emerge when several signals are separated from their superposition captured at distinct sensors. Despite the fact that rank estimation presents a significant challenge for modern NMF methods. This is comparable to the challenge of determining the eigenvalues of noise signals in a multichannel situation. A two-fold mistake occurs when each activation is carried out separately, making the deviation more susceptible to DNN approximation errors. As a result, the problem of the spectral overlay at the beginning of a dialogue or commotion is lessened and discriminative grounds are created. On the other hand, there are still several issues, including a lack of robustness and inadequate separation accuracy. Thus, in order to get around the aforementioned drawbacks, this research paper suggests an effective method for multichannel speech signal separation using a novel hybrid approach that combines enthalpy-based DNN (EDNN) and grasshopper optimization-based matrix factorization (GOMF). This method addresses all of the aforementioned drawbacks. In order to anticipate and eliminate unwanted noise from multichannel input signals, this research suggests a classification framework made up of STFT, GOMF-based rank estimation, signal eigenvalue identification, noise reduction, feature extraction, and classification. In addition, a unique hybrid technique that segments multi-channel speech signals using enthalpy-based direction of arrival (DOA), krill herd-based matrix factorization (KHMF), and SCNN estimation has been developed.

1.9 Database details

We utilised certain data sets available for free from the CHiME database [39] as the noise signal and the WHAM database [40] for the single channel speech separation. Additionally, we used voice and audio data from the TIMIT Corpus [41] and Noisex-92 Corpus, respectively. 10 phrases from TIMIT were delivered by 630 speakers from 8 distinct American dialect areas. Each of the 15 general sound kinds found in a normal setting on the NOISEX-92 lasts for roughly 4 minutes. Factory noise, F-16 noise, chatter noise, and other sounds may all be found on the NOISEX-92. To guarantee that the different components of each noise utterance are mixed with the clean speech utterances, we arbitrarily split each NOISEX-92 noise utterance into distinct pieces based on the temporal length of speech utterances. These sounds are primarily analogous to other common noises and are likewise transient.

In addition, several datasets that may be freely downloaded from the SASSEC07 database for the multi-channel speech separation. The four courses of source signals—four female voice sources, four male speech sources, three non-percussive music sources, and three music sources with drums—each with a 10 second period and 16 kHz investigation are used to construct the advancement information. In addition, we made advantage of 50 datasets of professional music recordings from SiSEC 2018. Here, the TIMIT corpus is used to choose

examples of clear language and diffuse noise. We utilised three real-time sound mixing mics from SiSEC 2011, level 3x5 (3 mixed signals–5 signal sources), and level 4x8 to examine the typical p-dimensional scenario (4 mixed signals-8 signal sources). many voices, both male and female. We combined 5 audio sources for the 3x5 example and 8 audio sources for the 4x8 example.

1.10. Hardware Tools

Speech separation is done in this research using MATLAB on a system with 6 GB of RAM and a 2.6 GHz Intel I-7 CPU. A programming environment for algorithm creation, data analysis, visualisation, and numerical calculation, MATLAB is the language of technical computing. It is the top maker of software for mathematical computation. Millions of engineers and scientists use MATLAB on a global scale to analyse and develop the technologies and systems that are altering our world. The matrix-based MATLAB language is the most natural language for communicating computer mathematics in the world. Data may be easily viewed and analysed thanks to built-in graphics. On a desktop environment, learning, exploring, and experimentation are simple.

1.11 Structure of The Thesis

The rest of this thesis is organized as follows:

Chapter 1 highlights the issues with speech separation on a single channel and many channels.

Chapter 2 gives a brief overview of the numerous problem-solving strategies that are frequently documented in the literature.

Chapter 3 Given that Single-Channel Source Separation is currently the most difficult challenge, this chapter proposes a revolutionary Time-Frequency non-negative matrix factorization and sigmoid base normalisation deep neural networks. According to the investigation's findings, the suggested method is preferable to the conventional strategy.

Chapter 4 In order to provide effective optimization-based speech separation, this chapter suggests an original RDNN-based speech/filter model (ERSS). The results of the investigation demonstrate that the proposed method is superior to the established one.

Chapter 5 In this chapter, we provide a technique for decoupling multichannel speech signals that combines Grasshopper Optimization-based Matrix Factorization (GOMF) with Enthalpy-based DNN. The experiment's findings will demonstrate that the technique we've suggested is preferable to established practises.

Chapter 6 The separation of sound sources with time-varying mixing quality induced by speech separation is an important research topic for enabling intelligent audio systems in real-world operating environments, and it is covered in this chapter. Multi-channel speech separation, enthalpy-based DOA, and score-based CNN estimation are all topics covered in this chapter (SCNN). Experimental results demonstrate the superiority of this approach over conventional practises. Compare the results to a number of proven subjects and algorithms, such as BP, CTF-MINT, CTF-MPDR, and Unproc techniques.

Chapter 7 Finally, a conclusion is offered, along with recommendations for further research.

Chapter 2

Literature review

2.1 Introduction

Speech is the primary means by which humans exchange information. Speaker and speech recognition is a common investigative technique that is only employed in everyday life. For many speech processing applications, speech separation serves as the fundamental framework. The system's performance suffers significantly when there are competing speaker signals present in the input mixture. The separation of mixed voice signals has been regarded as a significant and fundamental topic, with a wide range of applications in telecommunications, audio and speech signal processing, and medical signal processing. These are just two examples of the potential applications for audio and voice separation systems. Other applications include automatic speech recognition (ASR) in noisy environments and multimedia or music analysis, which purposefully combines information from multiple sources. Single-channel speech separation (SCSS) and multi-channel speech separation are the two types of speech separation systems (MCSS). A speech augmentation system is another name for SCSS. Source separation techniques can generally be categorized into two groups based on the acoustical configuration, the number of microphones, and the number of speakers: over-determined, where the number of microphones is higher than or at least equal to the number of unknown speakers, and under-determined, where it is lower.

This thesis' primary objective is to separate speech mixtures utilizing a single observation captured with a single microphone for both single channel and multi-channel speech separation. Research on multi-channel voice separation utilizing several microphones has already produced some astounding results.

Humans are capable of accurately and easily separating mixed signals, as seen by day-to-day existence. A machine cannot, however, perform such precise separation with ease. The single-channel separation problem is solved without the use of a reference signal, in contrast to the multi-microphone separation scenario.

2.2 Single-channel speech separation strategies

The key to the speech separation strategy is to model the process after the human separating mechanism, which is then mirrored in machine forms. The cocktail party effect describes how frequently in our daily lives we hear noises that are not isolated but rather combined with a noisy backdrop, such as traffic, crowds, radio, and television noise, depending on the surrounding circumstances. The target voice and the background noise can be distinguished by humans. However, as a system, it will be able to detect mixtures of various voice signals with varying time and frequency. Single channel speech separation is the process of isolating a particular needed speech signal from background noise or from a combination of speech signals when a single microphone is employed to record the speech mixture. One step in the speech separation process is single channel speech separation, often known as multiple input one output system (MISO). The SCSS problem can be solved using a variety of methods, including general signal processing approaches, computational auditory scene analysis (CASA) approaches, blind source separation (BSS) approaches, and model-based approaches.

General signal processing and CASA-based approaches are unsupervised approaches among the methodologies listed above because they look for features in the observation signal that can distinguish between speech signals and other signals. Contrarily, the BSS and model-based techniques are supervised approaches since they rely on sources' prior knowledge that was learned during a training phase.

Musicians frequently employed the harmonic model for single channel source separation. Michael Stark et al. (2011) [42] provide the Long Frame Associated Harmonic Model (LFAHM) to distinguish the two voice sources from a single channel. Through the use of harmonic frequency, this method solves the short time window overlapping issue. The pitch was estimated simply and precisely using the autocorrelation technique. Additionally, this method eliminates unvoiced portions from the mixture and surpasses the harmonic model in terms of SNR and quality. It produces improved accuracy in mixture separation and doesn't require any prior information of the speaker. However, this method cannot handle two or more unvoiced signals at once.

Monaural source separation was handled by Mohammadiha et al. (2013) [43] in a mixed signal containing voice and piano components. The energy of FM transmissions is determined using the discrete energy separation algorithm (DESA). A time-varying filter is developed in the time frequency domain to remove the interfering signal. In order to estimate the FM signal energy, instantaneous signal qualities that are limited in both time and frequency are used.

2.3. Deep learning techniques for single channel speech separation

DNN was utilised by Tae Gyoon Kang et al. [44] to map the data vector and related encoding vectors. Source separation, DNN training, and non-negative matrix factorization (NMF) training are the three stages of the suggested methodology. DNN-NMF performs better than earlier NMF-based approaches, although it is less adaptable.

Shuai Nie et al. [45] first proposed a DNN and Nonnegative matrix factorization (NMF) combination for speech separation. NMF first learns the spectra of the voice signal before reconstructing the signal and noise levels. The original speech content is preserved while the noise is removed using discriminative training with a scarcity constraint at a very low cost in terms of distortions and artefacts.

Time-varying masking is used to separate noise from speech input and handle channel mismatch. Once the system has been trained on clean data, A. Narayan et al. [46] suggested employing the diagonal feature discriminant linear regression (dFDLR) adaptation technique for the deep neural network and HMM for noise-resistant voice recognition. When dFDLR is trained on noisy log-Mel spectral characteristics, the best results are obtained. A number of scenarios, including clean, noisy, clean + channel mismatch, and noisy + channel mismatch, have been used to train the system. The system's flaw is that WER rises as a result of noise and channel mismatch.

For the deep neural network used for blind speech separation, Zhong-Qiu Wang et al. [47] recommended combining spatial and spectral data. A two-step Chimaera ++ network is used to analyse the temporal frequency dominance in order to determine the direction of the interested user. It works well with ASR that has several speakers. It performs poorly in environments with increased noise and reverberation, according to an experimental analysis of the RIR database.

Due to two major problems, the mixer's speaker counts and the speakers' positioning in relation to the target and masker speakers, speech separation is difficult. In order to address these problems, Yi Lue [48] investigated the use of the Deep Attractor Network (DANet) to project the time-frequency properties of mixed signals in high dimensional embedding space. The attractor (reference) point has a variety of effects on speaker clustering. Speech separation's permutation and speaker number issues are diminished by the attractor and permutation issues of DANet.

The speaker separation technique using text-independent speaker identification was examined by Nguyen Nang An et al. [49]. In order to learn speaker characteristics that can

handle variable length segments, CNN variations such as residual neural networks (ResNets) and visual geometry group (VGG) nets are used. CNN receives the log Mel's spectral properties. After the CNN layer, which generates input for subsequent layers of a predetermined length and concentrates on the discriminancy in the speaker characteristics, this structured self-attentive layer is used. The success of the system in a number of areas, including speaker authentication, speech emotion identification, and speaker diary, will be the next area of focus for this project.

For end-to-end time-domain speech separation, Y. Luo et al. [50] developed a completely convolutional time-domain audio separation network-based deep learning technique (Conv-TasNet). Conv-TasNet produces a representation of the speech waveform that may be used to identify individual speakers with a linear encoder. Several masks (weighting functions) are incorporated into the encoder output to separate the speakers. A linear decoder is then used to convert the updated encoder representations back into waveforms. By stacking one-dimensional dilated convolutional blocks, the building blocks of a temporal convolutional network can mimic the long-term relationships of the speech signal while yet having a small model size. With further advancements in its accuracy, speed, and computational cost, automated speech separation may one day become a standard and essential component of all speech processing technology.

A phase-sensitive goal function based on the signal-to-noise ratio (SNR) of the reconstructed signal was created by H. Erdogan et al. [51]. utilising a target function based on signal approximation. Performance has also been found to be enhanced by recurrent networks that are deeper and more dynamically accurate. Future prospects look good when language model information is more tightly integrated into speech separation and target phase prediction is used instead of the noisy phase.

A neural separation network with a clustering-based embedding was proposed by J. R. Hershey et al. Future research should focus on expanding training on datasets with a larger variety of audio formats and relevance to other fields like picture segmentation. In order to achieve end-to-end training for signal reconstruction quality for the first time, Y. Isik et al. [53] improved and expanded the deep clustering framework by extracting an embedding of spectrogram segments and estimating a mask for the separation component. The baseline system performance is first significantly enhanced by the authors by introducing better regularisation, a wider temporal context, and a more intricate architecture. This yields a 10.3 dB improvement in signal to distortion ratio (SDR) over the baseline of 6.0 dB for two-speaker separation and 7.1 dB for three-speaker separation.

A innovative technique for detaching a mixed audio sequence—a sequence in which several voices are speaking at once—was reported by E. Nachmani et al. in [54]. The novel approach makes use of gated neural networks that have been trained to split voices at various processing stages while preserving the speaker's stability in each output channel. The model with the greatest number of speakers is picked to estimate the actual number of speakers in a given sample, and each conceivable speaker count is represented by a distinct model.

The Hungarian approach, developed by Dovrat, S. et al. [55], substitutes the PIT loss and provides an ideal resolution to the permutation problem with a notably reduced temporal complexity, enabling the training of separation networks for a large number of speakers. Next, we offer a brand-new network architecture that uses stacked dilated convolutions before each pair of MulCat blocks. Even when writers make the permutation issues appear to be less challenging, complex designs still exist.

For speaker-independent multi-talker voice separation, Nasir Saleem et al. [56] presented a supervised binary classification strategy based on the DNN. In order to attempt larger efficiency gains of the proposed approach, we are committed to including the phase information in the upcoming work.

2.4 Separating multi channels of speech

A multi-channel speech separation is used to separate more than two voice signals sent over the same communication channel. According to past studies, the procedure is only deduced for two speaker. The algorithm may be expanded to accommodate many speakers. The between-cluster and within-cluster matrices can be expanded to accommodate many speakers. Iterative estimations can be used to calculate the energy ratios between different speakers. Quicker decoding approaches have been adopted because multi-talker settings become exponentially more difficult as the number of speakers increases.

In the current multi-channel speech separation technique, numerous microphones are employed to record the different speech signals. Hence Supervisory phrases are incorporated into the use of several microphones. However, if the speech combination is recorded using a single microphone, it is not possible to do so. The identical procedure taken unsupervised is always challenging. Several speech channels that were recorded in both a clean and noisy setting are separated from a single mixture using a hybrid vector quantization-based heuristic clustering algorithm (HVQHCA). Initial division of the input mixture into voiced and unvoiced speech fragments occurs by the algorithm. To extract different pitch values, spoken speech chunks are separated into segments. These pitch values are separated into numerous clusters for

diverse speakers using a dynamic clustering technique and the Silhouette value. Each individual's voiced segments, unvoiced segments, and complements of each voiced segment of the other individuals are all mixed into a single stream to generate the separated speech. The separations at the coarse and fine levels make the separated speech more accurate.

By using a simple pitch extraction technique for multi-speaker speech, Yi Luo et al. (2018) [57] revealed the potential for utilising the pitch information accessible from temporal processing for spectrum analysis. The fine weight function for the residual is derived by ascertaining the instants of desired and undesired speakers. The combined weight function of temporal processing is made by fusing the fine and gross weight functions. The degraded speech LP residual is multiplied by the combined weight function to get the enhanced residual. The time-varying all-pole filter made from the degraded speech is excited by the boosted residual to produce the temporally processed speech.

For the blind source separation of three speech samples in a real-world room environment, John R Hershey et al. (2016) [30] created a hybrid technique. Information-theoretic methods and the de-correlation technique are both used to provide superior source separation with quick convergence. The method is straightforward, computationally efficient, and intended for instant use. Additionally, no prior parameter estimation is required. It also used an innovative post-separation speech harmonic alignment to improve separated voice quality in a real-world situation. Minhas et al. concentrated on separation techniques for clear speech signals without considering background noise.

For speaker diarization, unsupervised speaker clustering has been proposed to divide similar voice segments into a number of speaker groups (Keisuke Kinoshita et al 2018) [57]. Prior to fine-tuning the segment borders throughout the re-segmentation process to obtain a final diarization hypothesis, the traditional techniques frequently performed speaker clustering on some initial segmentation. The author made use of the reference limits listed to determine the initial speech/non-speech boundaries. Using an iterative optimization strategy that alternates between clustering and re-segmentation until the diarylation hypothesis converges will yield better results if the number of speakers needs to be estimated. The iterative Variational Bayesian Expectation-Maximization-GMM clustering method finds a global optimum solution. The iterative optimization procedure recalculates the number of speakers using better-crafted and cleaner speech segments.

Y.-X. et al. (2013) [54] used a multi-channel system to enhance voice signals captured by scattered remote microphones in a car scenario. Each possible speaker in the car has a specialised directional microphone nearby that receives the corresponding voice signal. The

system creates an output signal in a hands-free phone conference call for a far-end communication partner and gets rid of the annoying crosstalk components from interfering sound sources that occur in multiple different mixed output signals because it's possible that other hands-free applications will be running concurrently. Through the use of a distinct signal processing block for interfering speaker cancellation, the crosstalk components of unwanted speech are eliminated. As part of the signal improvement, residual crosstalk and background noise are also diminished. Four speakers placed inside the automobile's interior and dynamically configured for a car setup are affected by a range of noise levels.

In order to combine BSS and noise suppression, Richard Lyon et al. (59) took use of the sources' sparseness in a short time frequency domain. A probabilistic model is used to improve the system by simultaneously suppressing the noise and separating the speakers in the event of active multiple speakers. This model takes into consideration the possibility of additive noise and captures the spatial information of the multi-channel recording. The estimation of source activity and the estimation of model parameters are the E and M steps, respectively, of the EM technique. The multi-channel adaptive filters are employed to remove noise and interference signals using spatial information.

2.5 Deep learning for multi-channel speech separation

Wang et al [60].'s deep clustering framework combined spectral and spatial data to better efficiently utilise the complementary spectrum and geographical information. By using phase difference features in the input, we can improve the predicted time-frequency masks by including both spatial and spectral information in the embeddings that deep clustering networks generate. Future studies will focus on combining the recommended approach with beamforming methods.

Compared to our earlier approach, Chen et al [61].'s innovative architecture for speech separation and multi-channel beamforming performs this combination more well. The suggested architecture is made up of a set of fixed beamformers, a beam prediction network, and a speech separation network created via permutation invariant training (PIT). The input beamformed audio signals are used by the beam prediction network to forecast the ideal beam for each speaker in the input mixture. PIT-based speech separation networks are presented in two different forms. We will eventually expand our research to include multi-talker voice recognition, and since it has been demonstrated that this method is more productive, we will jointly train each component more than once rather than just once.

A novel speech separation method was presented by Perotin et al. [62], and Perotin et al. [62] demonstrated the method's effectiveness using HOA materials. The calculation of a GEVD multichannel Wiener filter uses LSTM-based mask estimation. We want to assess the system's robustness to small inaccuracies in the projected DoAs in a subsequent investigation given the DoAs.

The deep learning-based multi-channel speaker separation technique developed by Wang et al.[63] makes use of both spectral and spatial data. The basic idea is to identify unique speakers using an augmentation network, allowing them to be distinguished from an approximated direction and with respect to specific spectral properties. To determine the speaker's direction of arrival, we only use the time-frequency (T-F) units that the target speaker dominates (DOAA two-channel permutation invariant training network that considers spectral and inter-channel phase patterns at the input feature level is used to assess the speaker dominance of each T-F unit. Tightly integrated beamforming, based on T-F masking, also makes use of the magnitudes and phase created. The combined training of the PIT and the augmentation network, the investigation of additional categories of spatial information, and a closer connection with beamforming techniques are just a few of the areas that could be the focus of future study.

A brand-new end-to-end mechanism for multi-channel speech separation was put forth by Gu et al. [64]. A proposed integrated neural architecture that separates speech into waveform-in and waveform-out components is the initial step. The traditional STFT and IPD are then reformulated by the authors as a function of time-domain convolution with a chosen custom kernel. Third, we made the fixed kernels learnable, allowing the architecture as a whole to be taught from scratch and fully data-driven.

Transform-average-concatenate (TAC), a straightforward technique for number-invariant multi-channel speech separation and end-to-end microphone permutation, was proposed by Luo et al. [65]. Before concatenating the output from the second stage with each of the output from the first stage and transferring it to a third sub-module, the ATAC module first translated each input channel feature using a sub-module, averaged the outputs, and then passed it to another sub-module. For each channel, the first and third submodules were shared. TAC can be considered a set-based function that is guaranteed to employ all of the data in the set while generating overall judgments, regardless of the permutation or quantity of set components.

Gu et alintegrated architecture.'s for learning spatial features directly from the multi-channel speech waveforms was developed in an end-to-end speech separation framework [66].

It is capable of learning useful spatial cues from the multi-channel speech waveforms in a completely data-driven manner. To achieve adaptive spatial filtering, this method makes use of time-domain filters that cover a variety of signal channels. These filters are built with the help of a 2D convolution (conv2D) layer, and the speech separation objective function is utilised to completely data-drivenly change their parameters. In order to calculate the inter-channel convolution differences (ICDs), we use a conv2d kernel that we created in part by drawing inspiration from the IPD formulation. It is anticipated that the ICDs will offer the spatial information necessary to distinguish between directed sources.

Gu et al. [67] introduced a unique multi-channel TSS framework that exactly imitates cRM estimation in the complex domain using a complex deep neural network (cDNN) with a U-Net topology. This framework was carefully constructed to make the most of temporal spectral-spatial data.

An ADL-MVDR framework that may be customised and used for multi-channel, multi-frame, and multi-channel multi-frame target speech separation tasks was proposed by Zhang et al. [68]. The proposed ADL-MVDR system solves the numerical instability problem that arises in traditional neural mask-based MVDR systems during cooperative training with neural networks by relying on RNN-predicted filtering weights. The ability of the suggested ADL-MVDR systems to generate practically any nonlinear distortions with minimal residual noise suggests the systems' aptitude to achieve the greatest objective scores (reflected by lowest WER). The disadvantage of the new ADL-MVDR system is that it needs more processing power than earlier, neural mask-based MVDR systems. Li et al[69] .'s three-step audio-visual multi-channel speech separation, dereverberation, and identification technique completely takes into account visual information. Future studies will improve the integration of the separation, dereverberation, and recognition components.

2.6 Methods for Blind Source Separation

There have been a number of time domain and frequency domain methods for single channel source separation presented. The time domain approach may encounter convergence problems and a significant workload if noisy chats are recorded in a noisy, busy environment. This is because numerous parameters must be examined. In contrast to the time domain, the frequency domain can simplify complex valued instantaneous blends for each frequency bin. It has substantially easier calculations and faster convergence than the time domain.

In order to retrieve the original source signals, a statistical method known as blind source separation searches for instantaneous mixes of a collection of source signals. The assumption behind BSS is that the mixing process must be linear. The BSS problem is frequently solved using independent component analysis (ICA) (Kevin et al. 2009) [70]; its extension necessitates that the sources be statistically independent of one another. The ICA approach to BSS generally seeks to invert the mixing process (de mixing) for recovering the original components by obtaining a linear transform of the mixes so that the recovered signals are as independent as possible.

To overcome the constraint of having a limited number of observations and to resolve the single channel source separation issue, some studies had used underdetermined BSS approaches (Benesty et al 2008) [38]. With these techniques, supplemental information is typically used to address the problem (such as a priori understanding of the statistical models of the sources).

Signal processing research has long been interested in the BSS problem. Strong principal component analysis and ICA are two examples of conventional BSS approaches. A potent BSS framework called non-negative matrix factorization divides data into activations and templates, or spectral templates and temporal activations for spectrograms (NMF). NMF presupposes that the data are not negative. Unsupervised BSS has drawn a lot of interest recently. Numerous unsupervised BSS techniques are covered in this section.

The variational auto-encoding-based single-channel blind source separation system developed by Neri et alnovel. [71] outperforms conventional techniques while automatically choosing the appropriate number of sources in data mixes. To disentangle (separate) data mixtures into low-dimensional latent source variables, a deep inference network is used. Each latent source is separated into its source signal by a deep generative network, whose sum matches to the input mixture. To automatically differentiate arbitrary-length films and universal audio waveforms, the proposed method has to be developed. In the lack of parallel clean data, Drude et altraining .'s technique [72] was offered as a way to train neural network-based source separation algorithms from scratch. It is expected to extend the current work to CHiME 5 challenge recordings in order to better synchronise authentic recordings.

The thorough method for blind source separation developed by Drude et al.[73] included probabilistic spatial mixture models, deep attractor networks, and neural deep clustering. The integration was accomplished by creating a mixed model that shared the same latent class affiliation variable between both modalities and had two separate observation distributions, one for the vector of microphone signals and the other for the embedding vectors generated by the

neural network. By integrating an extra speaker identification embedding, Haeb-Umbach et al. [74] presented a deep attractor network for blind source separation and speaker re-identification.

The new single-channel blind source separation (SCBSS) algorithm was created by him and his co-workers [75] and is based on multi-channel mapping and Independent Component Analysis (ICA). It assumes that mixed signals originate from dynamic systems in which each component is impacted by interactions with other components and signals are instantly mixed in a linear fashion. The authors state that the algorithm will be enhanced in the future to achieve online SCBSS in accordance with the dynamic system concept.

A technique for bootstrapping a single-channel deep network for source separation that was inspired by biology was disclosed by Seetharaman et al. in [76]. To train the model, noisy separation estimates from stereo mixes are created using a spatial audio source separation technique. Even when the method that taught it didn't give it the necessary cue, the trained model can recognise sources in single-channel mixes. The authors created a confidence metric for the spatial method's output. Any clustering-based separation technique may specify a comparable confidence metric to lessen the effect of subpar training estimations on model training.

The two stages of the process are the training phase and the testing phase. Using sparse coding, nonnegative matrix factorization (NMF), or ICA, the voice sources are projected onto a set of fundamental operations during the training phase. Makino et al. (2007); Cherry et al. (2003). (2007) (Benesty et al.) [77]. During the testing phase, the necessary speech signal is separated from the speech mixture by comparing the speech signal to the statistical model created during the training phase. Statistical model-based methods estimate the clean speech spectrum in noisy environments using a statistical estimation framework. The techniques employ maximum likelihood, least mean square error, and a posteriori estimator as well as other probabilistic-based speech spectrum estimators.

Supervised NMF generates new sources by combining sources from a learned set of bases for each source in the mixture. Hyvarinen et al. (1999) [78] combined the model-driven separation technique with ideas from sparse coding and NMF by choosing the appropriate number of bases in the training. In conventional NMF, which ignores phase information, the spectrogram matrix of the mixed signal is factored into the sum of rank-one source spectrograms. In Virtanen, the presumption that phase shouldn't be factorised and its consequences on separation are investigated [79]. If the underlying source spectrograms are

given a priori, there is an improvement over NMF that tracks the distribution of the spectrogram points of the mixture.

In practise, there are no individual source recordings available. Yi Luo et al. (2019) [80] proposed an NMF method for monaural blind source separation to solve this problem by using mixed audio recordings to train the source models. In a single-channel scenario, speech mixture has been separated using vector quantization (VQ) and NMF. Despite the apparent differences between the two methods, the VQ strategy for model-driven separation is remarkably similar to the supervised NMF separation strategy.

According to Hennequin et al. (2020) [81], specific features affect a voice application's accuracy more significantly than specific generative models do. A sub band perceptually weighted transformation (SPWT) was applied to the magnitude spectrum to improve the performance of a single-channel separation scenario. The author specifically contrasted the SPWT, magnitude spectrum, and log-spectrum feature types. A rigorous statistical analysis is used to evaluate the efficiency of a VQ-based SCSS framework in terms of the lowest error bound. Two trained codebooks that were utilised to conduct the primary separation evaluation on the quantized feature vector of speakers form the basis of this methodology. The simulation results show that the transformation offers a viable option for improving the separation efficiency of model-based SCSS. It is also mentioned that it generates a higher spectrum SNR and a lower-error bound for spectral distortion when compared to other characteristics.

Table 2.1. Summary of Speech Separation Methods

Author and Year	**Methodology**	**Database**	**Evaluation Metrics**	**Application**
Chang et al. (2015) [82]	Deep Neural Network – Non-negative matrix factorization (NMF)	TIMIT and NOISEX-92 noise dataset.	SDR -8.74 SIR -11.20 SAR -13.91 PESQ -2.23	Source separation, speech enhancement

Stephan et al. 9 (2018) [83]	combination of DNN and Nonnegative matrix factorization	TIMIT and NOISEX-92 dataset	SDR -9.8 SIR-14.7 SAR-10.2 PESQ- 0.59	Speech separation
Wichern et al. (2019) [84]	Diagonal feature discriminant linear regression (dFDLR) and Deep Neural network (DNN)	Aurora-4 medium--large vocabulary	Word Error Rate - 4.8% (Clean Training)	speech separation and noisy speech recognition
Shi et al (2019) [85]	combination of spatial and spectral features for deep neural network	WSJ0-2MIX using up to two microphones, WSJ0-3MIX using up to two microphones	SDR- 10.4 for WSJ0-2MIX SDR-7.9 for WSJ0-3MIX	blind speech separation
Liu et al. (2020) [86]	Deep Attractor Network (DANet)	Wall Street Journal dataset	SDR – 10.4 (2 speaker) SDR- 8.5 (3 speaker)	Speech separation
NGUYEN NANG AN et al. (2019) [87]	Convolutional Neural Network	VoxCeleb database	Accuracy-88.2% (VGG+ Self attention Layer) Accuracy-90.8% (ResNet+ Self attention layer)	Speaker identification
Y. Luo (2020) [88]	fully convolutional time-domain audio	WSJ0-2mix and WSJ0-3mix datasets	PESQ-3.24 (WSJ0-2mix) and 2.61 (WSJ0- 3mix datasets)	Speech Separation

	separation network (Conv-TasNet)			
Wang (2018) [89]	LSTM neural network with a phase sensitive loss function	CHiME-2	SDR-14.75 SIR-20.46	speech separation
Luo (2019) [90]	deep clustering	WSJ0 corpus	SDR-6.8 (2 Speakers)	Speech separation
Shi et al (2020) [91]	To accomplish end-to-end training for signal reconstruction quality, the deep clustering framework was extended.	WSJ0 corpus	SDR-10.5(2 Speaker) SDR-7.1 (3 speaker)	Single-Channel Multi-Speaker Separation
Han et al (2020) [89]	two bi-directional RNNs and a skip connection are combined in a new recurrent block.	WSJ-mix dataset was extended to include mixtures of 5 speakers	SDR-20.12 dB (2 speaker) SDR- 10.6 dB (5 speaker)	Voice Separation with an Unknown Number of Multiple Speakers
Fan, S et al [2020].[90]	training for permutation invariance with the Hungarian method	WSJ-5mix Libri-5Mix Libri-10Mix Libri-15Mix Libri-20Mix	SDR-12.72(WSJ-5mix) -7.78 (Libri-5Mix) -5.66 (Libri-10Mix)	Many-Speakers Single Channel Speech Separation

			-4.26 (Libri-15Mix) -13.22(Libri-20Mix)	
Yi Luo et al (2018) [91]	Fully connected Deep neural networks based binary classification	720 IEEE speech utterances WSJ0-2mix	PESQ- 2.84(2 talker), 2.7(3 talker), 2.57(4 talker) SNR-6.85(2-talker), 2.7(3 talker), 2.57(4 talker)	single channel speaker independent multi-talker speech separation
Yi Luo et al [2018] [92]	Utilises a deep clustering architecture that integrates spectral and spatial characteristics.	wsj0-2mix dataset	SDR- 12.9	Multi-Channel Speaker-Independent speech separation
Nima et al (2019) [94]	Fixed beam formers with Bi-LSTM integration	anechoic speech signals, internal collection of utterances spoken by 44 speakers, WSJ SI-284	SDRs of different separation systems for different mixing conditions.	multi-channel far-field speech separation
Keisuke Kinoshita et al [2018] [70]	Recurrent neural networks	Ester dataset	word error rate-11%	Multichannel speech separation
Naoya Takahashi et al [2019] [71]	combines spatial and spectral data	wsj0-2mix corpus.	SDR-10.9	Multi-Channel Speaker Separation

	for deep learning.			
Vincent et al [2003] [72]	end-to-end approach	WSJ0 2-mix	SI-SNR- 11.6	multichannel speech separation.
Hao et al [2020] [95]	transform-average-concatenate	Libri speech dataset	Si-SNR- 12 for 6 mics	number invariant multi-channel speech separation
P.-S. Huang, et al [2014] [97]	End-to-end speech separation framework integrated architecture for learning spatial information directly from the multi-channel speech waveforms.	WSJ0 2-mix dataset	SI-SDR- 11.9 SDR-12.3	Multi-Channel Speech Separation
Felix Weningeret al [2021][96]	U-Net structure is used to carefully create the complex deep neural network (cDNN).	Original speech data is collected from You tube	SI-SDR- 12 WER- 17.03	Multi-channel Target Speech Separation in Complex Domain
Yusuf Isik et al [2018] [61]	ADL-MVDR framework	Mandarin audio-visual corpus	PESQ-3.46 SI-SNR-15.43 SDR-16.03 STOI-93.7	multi-channel multi-frame Speech Separation

			WER-12.31	
Dong et al [2017] [65]	DNN-WPE and spectral mapping	LRS2 dataset	PESQ-2.49 SRMR-8.71 WER-22.38	Audio-visual multi-channel speech separation, Dereverberation and Recognition
Neri et al [2021]	variational auto-encoding	MNIST and MUMS dataset	SI-SDR-17.10 SIR-29.55 SAR-18.20	unsupervised single-channel blind source separation
Yi Luo et al [2018] [64]	unsupervised spatial clustering algorithm	WSJ sets	SDR-9.5 PESQ-0.40 STOI-0.18 WER-29.3	Multichannel Blind Source Separation
Drude et al [2019]	Deep clustering, deep attractor networks, and probabilistic spatial mixture models are used in an integrated method for blind source separation.	Wall Street Journal sets	SDR-6.8 PESQ-0.60 STOI-0.15 WER-33.4	acoustic blind source separation
Morten Kolbæk et al [2017] [66]	Deep Attractor Network based system	Wall Street Journal sets	SDR-9.2 SIR-16.4 SAR-10.6	Speaker re-identification and blind source separation
He et al [2018]	multi-channel mapping	TIMIT dataset	SNR for different sampling points	single channel blind source separation

Seetharaman et al [2019]	To train a deep learning source separation model, stereo mixes are subjected to unsupervised spatial source separation that results in the first breakdown of the mixtures.	wsj0-2mix	SDR-2.9 SIR-13.5 SAR-3.7	Single-channel source separation

2.7 Research Gap

In this investigation, distinctive execution measures to appraise the word mistake paces of reproduced behind-the-ear listening device flags and identify the azimuth point of the objective source in 180-degree spatial scenes. These measures get from phoneme back probabilities created by a profound neural organization acoustic model.

- In existing NMF, rank estimation is a major issue. This is equivalent to the issue of recognizing noise signal eigenvalues in a multi-channel environment.
- In existing Deep Neural Networks (DNNs), assessment of the activations is acted in a different lead to a twofold error issue and create the departure progressively delicate to approximation inaccuracies of DNN.
- This makes discriminative bases and reduces the issue of spectral overlay in the beginnings of discourse and commotion.
- In another hand, numerous issues remain to comprise poor separation accuracy and absence of robustness.
- In a previous study, the hybrid Grasshopper Optimization-based Matrix Factorization (GOMF) algorithm shows great potential in the Multichannel speech separation. However, GOMF has a parameter initialization problem and leading to poor separation performance.

- Instead, a joint creation of the GOMF model parameter approximation and source localization delinquent.

In this section, some existing techniques and their drawbacks are discussed. To overcome all these research gaps, methods such as hybrid GOMF and Enthalpy based Deep Neural Networks have been proposed in the present research work.

2.8 Summary

As various methods of speech separation and classification for Single and multi-channel speech signals exists, it is difficult to draw meaningful conclusions about the merits of anyone approach over another. The techniques developed in this thesis are useful as this leads to improve the SNR advantage of signal separation and classification is carried out in real-time. Several supervised separation systems have been proposed. The in-depth learning methods used for supervised speech separation increased the rate of progress and increased the separation efficiency. Also, reliable assessment of time-frequency masks from the conversation is challenging, especially when there is room echo in the mix. From this survey, to overcome this issue, we propose an efficient optimal reconstruction-based speech separation (ERSS) using a hybrid deep learning technique. The upcoming chapter will explain based on single channel source separation using FTNMF and softmax.

Chapter 3

Time-Frequency Non-Negative Matrix Factorization (TFNMF) and Sigmoid Base Normalization Deep Neural Networks for Single Channel Source Separation

3.1 Introduction

For single channel source separation problem, multiple clean speech signals data sets have been considered for investigations. The work has been carried out using Time- Frequency Non-Negative matrix factorization (TFNMF) and Sigmoid Base Normalization Deep Neural Networks (SNDNN). The human auditory system can, with some concentration, hear the speech of a specific speaker in such a situation. It implies that the human hearing system is capable of distinguishing between different sources and efficiently identifying the source of interest. However, the human auditory system also has significant limitations in terms of the perception of the incoming information. Not only are researchers interested in creating an effective speech recognition system that mimics human auditory function, but they are also interested in extracting more information from the input signal than a human can.

When the number of data collecting sensors (microphones) is restricted, the situation for the machine gets complex. For instance, a cell phone, which is one of the most prevalent electronic devices that people carry around with them during the day, contains just one microphone. When a user speaks a command into their mobile device, it must be able to identify it even in challenging circumstances. Such situations may arise when the target signal is mixed with background noise (such as the chatter of a train, car, or factory), the speech of another speaker, or music. Monaural recording is the practice of recording many sources simultaneously using a single sensor (microphone).

The problem of monaural speech separation, also known as single channel speech separation, is the main focus of this chapter. This is followed by the recognition of the target speech. The challenge is to distinguish individual speech signals from the mixture with an unknown mixing pattern using a monaural signal because, in a practical case, the level of mixing is also not specified.

The two stages listed below make up this chapter. The training phase comes first, and the testing phase follows. The testing stage employs a single-channel multi-talker input signal, whereas the training stage uses a single-channel clean input signal. This distinction between the

two testing and training phases allows for more accurate comparisons. The input signals from these testing and training phases are sent to the short-term fourier transform (STFT). When extracting features from spectrograms created by STFT, which transforms input clean signal into spectrograms, TFNMF is the approach used.

3.2 Monaural Source Separation

The problem of monaural speech separation, also known as single channel speech separation. Two fundamental approaches can be taken to solve the single-channel mixed speech recognition challenge. First, the mixed signal must be separated, and then the separated signals must be recognized. To recognize clear speech, a variety of effective speech recognition models are available. Therefore, this thesis' primary goal is to roughly approximate each individual speech signal from various types of mixed speech signals. According to the desired result, the source separation issue can be divided into two categories: "target versus all" and "audio modification."

It is common knowledge that a signal is also a concoction of various independent components. A combination of n sources will therefore contain N numbers of distinct components, where N > n. When separating the target source from the mixture, the "target versus all" problem was used. The isolation of all mixing components is necessary to solve this issue. one of the main tasks of the mixed signal. Together, various mixing elements created the goal signal. An easy solution to the "audio modification" problem can be found by recombining various mixing components of various separate signals.

The separation of a singing voice from any musical tune illustrates the applicability of the "target versus all" dilemma. Another illustration of the "target versus all" challenge is the speech identification of the target speaker in a loud environment and the separation of the individual signals of other speakers in a cocktail party situation. The audio editing can be used for current audio remixing, hearing aid signal augmentation, etc.

3.3 Time Frequency Non-Negative Matrix Factorization for Source Separation

The extraction of the signal's mixing components serves as the first step in the monaural source separation process. Speech signals are frequently processed in the time-frequency domain. Short-term Fourier transforms (STFT) of speech are thus employed during the mixing component extraction procedure. The enormous dimensionality of the Euclidean space in which STFT's time-frequency data is embedded is a characteristic of this technology [6]. It is

necessary to minimise the high dimensionality of the input data in order to separate individual signals from mixed signals. Additionally crucial is raising the standard of data analysis. PCA, SVD,TFNMF, and others are some of the common methods for dimensionality reduction. These methods can also be applied to the source separation task.

A mixed signal needs to be separated in many different situations, such as when speech is mixed with music, noise, or another voice at various decibel levels and in various acoustic environments. When recording speech with a single microphone or a group of microphones, the quantity of input mixed signals may also vary in various situations. Any of the previously suggested source separation methods may not be the best option in all circumstances. Different source separation strategies have been demonstrated to be applicable for a variety of mixed speech separation issues by researchers.

Due to its ability to represent data in a non-negative manner, NMF has quickly become one of the most popular source separation approaches. Like the pixel intensity of an image or the spectrogram of audio, many sample data points of a signal are non-negative in nature. It is anticipated that dimensionality reduction algorithms would show this data in a non-negative way. PCA and independent component analysis (ICA) cannot ensure non-negativity in such a circumstance. The existence of basis vectors is also shown by the presence of non-negative components. This provides inspiration to develop a non-negative decomposition of data solution.

In the past, NMF has the potential to be used in a variety of applications where non-negativity is a crucial requirement, such as picture enhancement, text clustering, and speech separation. The ability to modify the decomposition process in accordance with the application and other criteria, such as orthogonality, sparseness, uniqueness, etc., is what makes NMF so appealing. In a matrix X, where each column X I represents an observation, such as a picture, a spectrogram, or probability, observations of any signal are generally accumulated.

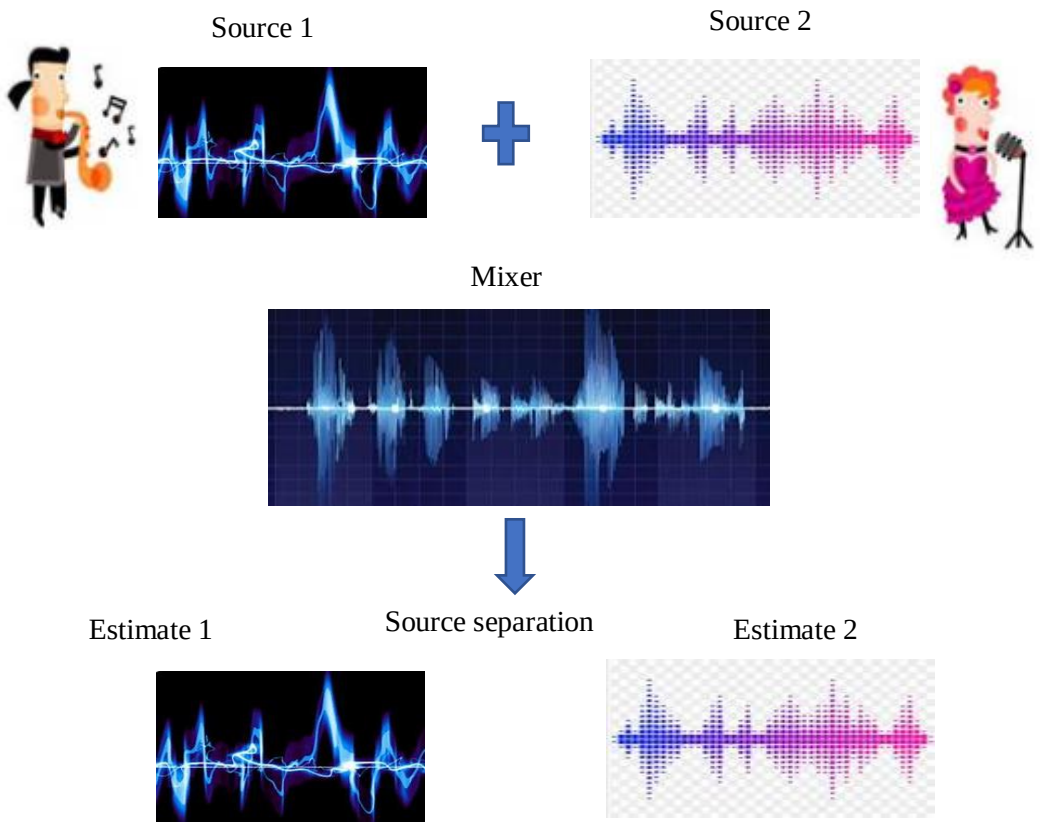


Figure 3.1: Single channel speech separation

A disruptive speaker, like the one in the illustration, or any other noise maker could be the second source. In this scenario, a listener may want to focus on a single target speech signal or on both signals separately. If noise is blending into the target signal, the noise may be isolated or suppressed; however, if the mixing signal is speech, both signals should be clearly separated, as seen in Figure 3.1.

As various speakers are speaking at roughly the same time, the goal of speech partition is to emphasise each speaker's mixed language speech. Multiple sources are separated from a single channel using SCSS (Single Channel Speech Separation). Automated speech recognition (ASR), hearing aids, and speaker recognition are just a few of the applications. Traditional single-channel speech separation approaches include computer auditory scene analysis (CASA) and TFNMF. To imitate sound processing by the human hearing system, CASA employs certain organising principles and appropriate decoupling signals.

Pre-learning is an essential consideration of the classification and regression process in digital signal processing. To minimize the overall design cost, these learning methods integrate the concept of a data matrix. The Time-Frequency NMF Non-Negative Matrix Coupling

(TFNMF), the most generally used pre-learning technology, is one of the most widely utilised technologies. Sound signal detection, environmental impact on speech recognition systems, and numerous functionalities of sound sources are all covered by TFNMF technology. Background interference from the primary target requirements for speech separation [98]. It's a signal processing feature that's useful in a variety of applications, including mobile communications, audio prosthetics, accurate speech, and speaker recognition. The human auditory system has a significant ability to distinguish one sound source from multiple others in a mixed environment.

Non-negative matrix data X is generated through the TFNMF approximation technique. The spiral TFNMF crucial attenuation is one of the most important instruments for signal processing and machine learning. TFNMF is the most effective and efficient way to distort fundamentals, and it offers a number of advantages over environmental resource separation. The basic goal of voice separation in a single microphone recording is to remove background noise from the target speaker. The solution covers the fundamentals of individual hybrid signals, from mixed signs to temporal frequencies, which are employed in a variety of applications including voice communication, speech coding, and authentic speaker learning methods. People who compete with various sound signal sources and background speakers in good complex surroundings focus on the auditory interest in signal combinations of complex signals, and humans excel at solving issues, according to the Cocktail Party Problem. Hearing-impaired audiences had more trouble with all interface speakers and intermediate spatial reversals than ordinary hearing aids, according to studies. Music recovery's major purpose is to assess and rely on the sound and background of music in advanced apps that contain information on reusing music recovery. To this purpose, a supervised technique, particularly one based on in-depth research, should yield current outcomes.

It would be used for both undeclared speech separation and correct speech extraction, which will improve application and usefulness. To facilitate utility extraction, additional speaker recognition steps can be built to identify target speakers from publications on undeclared segregation networks. Both approaches have benefits, and structural abstraction skills for uninformed speech extraction and undeclared speech separation are desirable. Allow extraction from multiple speaker outputs to identify target speakers. The most fundamental technique of determining filters is based on the time-frequency (TF) coverage, which determines how the TF mask is formed. No limit values are generated because this limit is appropriate and yields a modest approximate error (0.36 dB in Oracle tests). Monorail sound requires a single recorded microphone device to distinguish the target speaker from the background speaker. Methods of speech recognition Automatic Speech Recognition (ASR) is

crucial in the development of hearing aids. Finally, we detected the noise mistake in speech using the softmax classifier and removed it using the proposed technique.

3.4 Proposed system for TFNMF based DNN softmax

The proposed DNN softmax system based on the TFNMF is depicted in figure 3.2. There are two parts in it, including training and testing. Following the training phase is the testing phase. A single-channel multi-talker input signal is used in the testing phase, but a single-channel clean input signal is used in the training phase. Between these two testing and training phases, this is the primary distinction. Input signals for the Short-Term Fourier Transform come from both the training and testing phases (STFT). These concepts are equally thorough and insightful as the sections that follow;

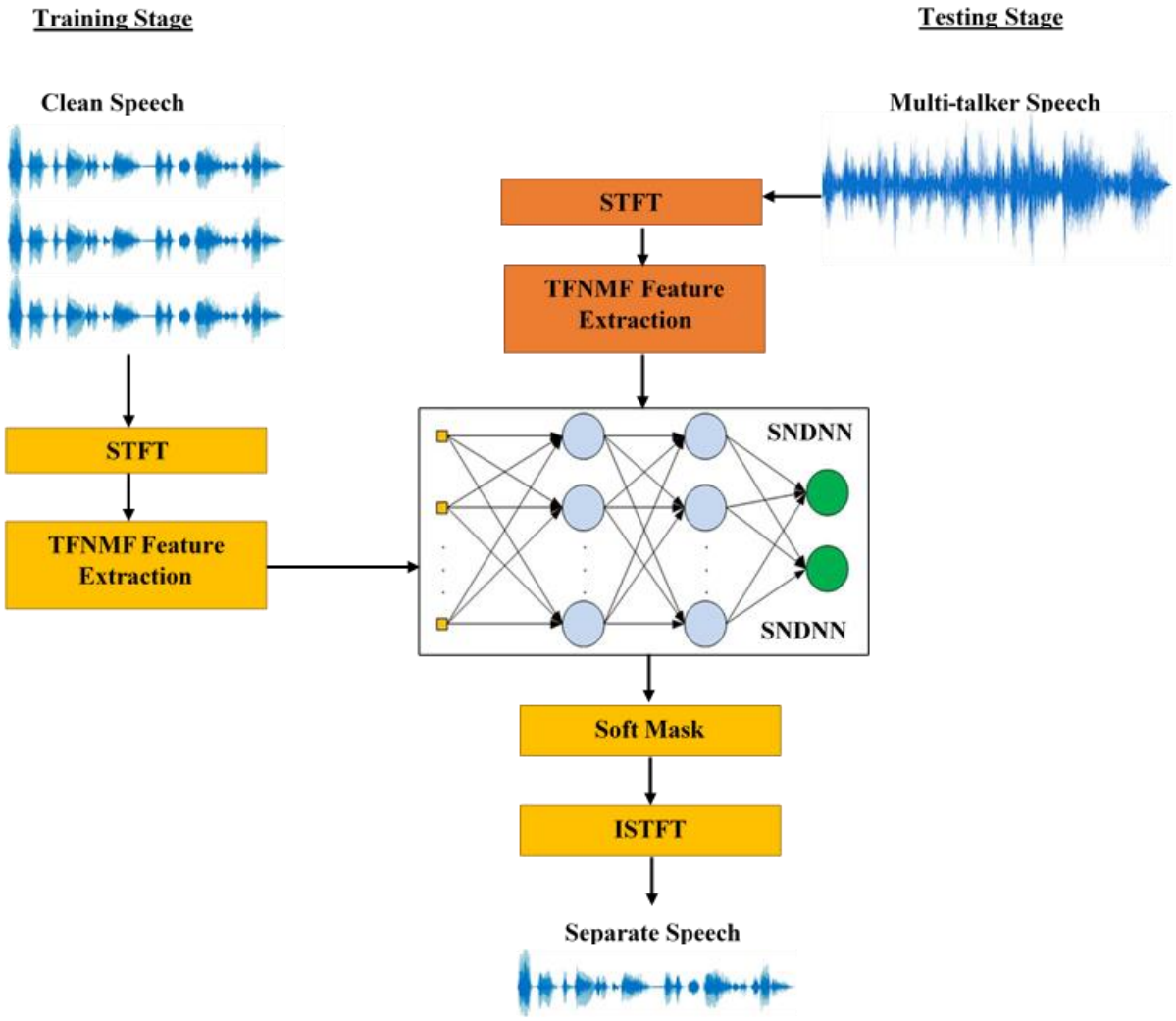


Figure 3.2: Block diagram illustrating the suggested approach

3.5 Algorithm of Training Stage:

Input: Mixed speech signal

Output: Signal channel speech signal

Step 1: Initially take the mixed speech signal from the database.

Step 2: Input signals pass to STFT.

Step :3 Spectrograms are produced by STFT from input clean signals.

STFT is a typical Fourier transform decomposition if the sign fluctuates over time or is ambiguous.

$$Z(y,\ f) = \int z(y_1).h^*(y_1 - y).e^{-(i2\Pi fy)}dy_1 \tag{3.1}$$

Comparing the spectrogram to the conventional Fourier change and range, the following design is possible:

$$U_z(y,\ f) = \left|Z(y,\ f)\right|^2 \tag{3.2}$$

It is typically employed to examine signals that evolve over time. The spectrogram breaks down the sign into a number of smaller components and calculates the range of each component, letting us know where various frequencies converge. a device that converts single-channel mixed sounds into intricate spectrograms. Then, using Time-Frequency non-negative matrix factorization, characteristics are obtained (TFNMF).

Step: 4 Feature Extraction based on TFNMF,

Here, Cohen's class of temporal frequencies for signal has a discrete-time form.

$$Z_{uu}(t,\ f) = \sum_{m=-\infty}^{\infty}\sum_{n=-\infty}^{\infty}\phi(n,m)u(t+n+m)\times u^*(t+n-m)e^{-j4\pi fn} \tag{3.3}$$

The time index and frequency index are denoted everywhere by the letters t and f, respectively.

Step :5 The distribution's kernel, which is dependent on both the time and lag factors, defines the distribution.

Step :6 In order to calculate the cross-TF between two signals,

$$Z_{u1,u2}(t,\ f) = \sum_{m=-\infty}^{\infty}\sum_{n=-\infty}^{\infty}\phi(n,m)u_1(t+n+m)\times u_2^*(t+n-m)e^{-j4\pi fm} \tag{3.4}$$

Step : 7 Expressions 3 and 4 currently explain the succeeding data spatial t-f distribution (STFD) matrix.

$$Z_{uu}(t,f)=\sum_{m=-\infty}^{\infty}\sum_{n=-\infty}^{\infty}\phi(n,m)u(t+n+m)\times uu^{*}(t+n-m)e^{-j4\pi fm} \quad (3.5)$$

Anywhere $\left[Z_{uu}(t,f)\right]_{i,j}=Z_{u_i,u_j}(t,f), for\, i,j=1...n.$

Step :8 The STFD matrix can be defined generally as

$$Z_{uu}(t,f)=\sum_{m=-\infty}^{\infty}\sum_{n=-\infty}^{\infty}\phi(n,m)\Theta u(t+n+m)\times u^{*}(t+n-m)e^{-j4\pi fm} \quad (3.6)$$

Anywhere is the kernel connected to a few speech signal data that represents the Hadamard product. Features are recovered based on non-matrix factorization after time-frequency estimation.

Step :9 In addition to using a non-negative grid with a focus on binary nonnegative lattices, NMF processes regressions.

$$UV\approx XY \quad (3.7)$$

Step :10 Whenever in the domain of K sections and N lines with nonnegative components. Following that, the NMF prototype can be coupled to a noise matrix in the following ways:

$$UV=XY+E \quad (3.8)$$

Step :11 calculations for measuring the matrices X and Y as the objective matrix UV and solving the NMF [41] problem. They involve replacing the following valuation requirements for each lattice:

$$\begin{aligned} X &\leftarrow \arg\min_{X} C(UV \parallel XY) \\ Y &\leftarrow ar\min_{Y} C(UV \parallel XY) \end{aligned} \quad (3.9)$$

Step :12 Due to the constraints, each component of matrix W and the component in the kth row and rth column of matrix X serve as a measure of how far apart matrices A and B are from one another. Using well-known "distance" measurements like the Gulbach-Leipler difference and the Euclidean distance, it is possible to analyse C (V || XY). In terms of Euclidean distance, C (V || XY) is precisely defined as follows:

$$C(UV \parallel XY)=\frac{1}{2}\parallel UV-(XY)\parallel_F^2 \quad (3.10)$$

$$\begin{aligned} X &\leftarrow X\otimes[(UVY^{T})\phi(XYY^{T})], \\ Y &\leftarrow Y\otimes[(X^{T}UV)\phi(X^{T}XY)], \end{aligned} \quad (3.11)$$

Wherever $\otimes and\ \phi$ term element-wise multiplication also division.

3.6. Testing stage:

This phase's testing is repeated using the same methodology. Inputs are treated as multi-talker speech in the testing phase as opposed to clean speech in the training phase. Characteristics from the training and testing periods are finally retrieved using a Time-Frequency non-negative matrix factorization. The SNDNN classification algorithm, which is described below, is used to categorise all features after feature extraction using the SNDNN classification step.

3.6.1 Classification algorithm using SNDNN

Our work makes use of a novel development based on sigmoid normalisation (SN) combined with DNN in place of the many existing procedures that are just based on DNN. One of the cutting-edge techniques we study is this one. Data is initially passed to the DNN convolutional layer, which uses the sigmoid to normalise the data. Once at the convolutional level, the layer progresses to the highest grouping level before repeating the process to reach the highest retrieval level.

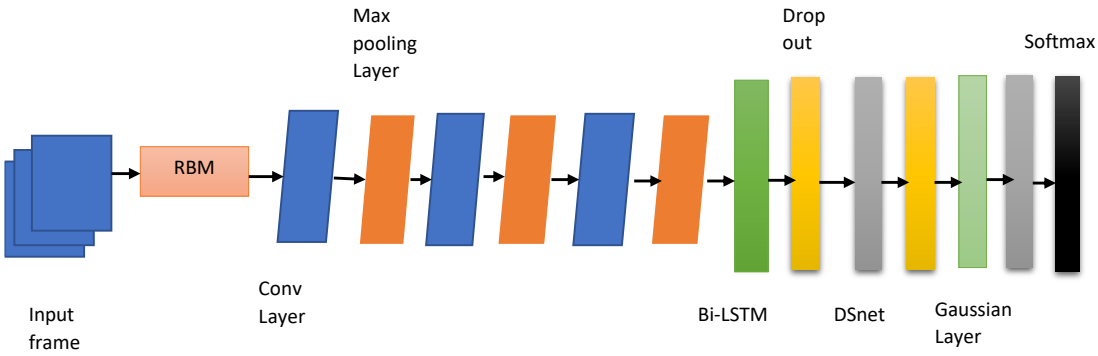


Figure 3.3: Reconfigured softmax-CNN Architecture

Data-related information is contained in the convolutional layer. The Softmax regression is connected to this layer. The basis of SN's initial processing is pure voice and input from multiple speakers. The output signal is sent to the highest grouping level once the procedure is complete, which minimises the display of limit values and calculations.

Algorithm for Softmax

Step :1 Define data

data = [speech database]

Step :2 Calculate softmax

result = softmax(data)

report the probabilities

Step :3 Calculate the softmax of a vector

def softmax(vector):

e = exp(vector)

return e / e.sum()

print(result)

Step :4 report the sum of the probabilities

print(sum(result))

A. Convolution layer

A matrix or kernel-based initial layer of the network is used to recover the signal from the first clear format. The signal's qualities are maintained in relation to one another via understanding the signal. Pay close attention to the spectrogram fields to comprehend the following degrees of work folding. Equation (3. 12)'s criteria are met by this layer, and the result of the adjustment is what is referred to as element mapping in each instance.

$$y_k = \sum_{n=0}^{N-1} x_n h_{k-n} \tag{3.12}$$

Wherever the input qualities are, there are filters and a number of necessities. Its output is the yield vector. The elements of the vector are represented by the subscripts.

B. Normalization layer

Sigmoid-based normalisation, the network's next layer, merely points to a comparable area or known route. To do this, a signal is normalised, which lowers its distortion to a constant mean of 0. The single-channel source separation's range is expanded by sigmoid-based normalisation (3.13), (3.14).

$$SG(u) = \frac{1}{1+m^{-u}} \tag{3.13}$$

$$SGBN = SG(u) \times \frac{X}{X_{max} - X_{min}} \tag{3.14}$$

In Single-channel source separation X, the lowest and maximum values are shown by SG (u), the sigmoid based normalised or Single channel source separation, the Euler's number.

C. Max-Pool layer

The pooling layer seeks to streamline the system's computation and boundary value presentation. When the max-pooling layer is employed to minimise the dimensionality of the signal, it has an effect on both the next level and the strength of the neurons. This also goes by the moniker of the downsampling layer.

D. Completely connected layer

Since there are a large number of neurons in the preceding layer, each neuron receives information from that layer. Softmax is a representational tool that can handle multiple classes because the name of the logistic regression is "0, 1".

$$p_i = \frac{e^{x_i}}{\sum_1^k e^{x_i}} \tag{3.15}$$

Wherever the network's input is based on entropy and the outcomes of recognising the harm or otherwise normal, SNDNN is utilised to classify it. We sequentially implemented the entropy-cantered deep neural architecture. In its boundary learning, it also involves mutual pre-preparing and modification stages.

E. Training Stage

Step 1: We display the visible units' advice for the training vector's selected features.

$$E(x,y) = -\sum_{i=1}^{I}\sum_{j=1}^{J} Q_{ij} f_{si} y_j - \sum_{i=1}^{I} \alpha_i f_{si} - \sum_{j=1}^{J} \beta_j y_j \tag{3.16}$$

Wherever it indicates the symmetric association duration between the visible constituent and the concealed component, the term "predisposition" denotes the number of understandable and hidden processes [99]. The position vector's logarithmic likelihood with respect to the weight is satisfied by the major conflict. In an RBM, covered units are not immediately impacted, but it is rather straightforward to make a case for

$$\rho\left(y_j = 1 \mid f_{si}\right) = \zeta\left(\sum_{i=1}^{I} Q_{ij} f_{si} + \alpha_j\right) \tag{3.17}$$

Anywhere $\zeta(x)$signifies the strategic sigmoid capacity $\frac{1}{(1+\exp(x))}$, f_{si}, h_j denotes the unbiased sample.

Step 2: We match the supplied hidden and explicit units to the evident and invisible units. The sharpest random rise in the log-likelihood of location data might be produced, according to this, by more direct learning principles, such as the ones that follow:

F. Fine-tuning phase

In essence, the fine-tuning stage is a normal back spread computation. System implementation organisations frequently achieve performance levels above the SNDNN. The reconciliation record is additionally output while the extra weight is processed or reviewed. In this situation, the SNDNN classifier is crucial because it will collect the required information and utilise it to finish the process, allowing it to position results that could lack certain qualities.

3.6.2. Inverse STFT (iSTFT) operation:

The extracted output signal can then be subjected to the reverse STFT technique to change the voice signal's final channel interval. Finally, get the split extracted signal.

3.7. Dataset Description:

We used several data sets from the WHAM database and the CHiME database that were freely available for use in the investigation. From the CHiME database, the current data set was chosen as the noise signal. The SIR, SDR, PESQ, and short-term objective understanding (SOUE) indicators are employed by the system to assess the objective indicators (STOI). We are analysing the offer to gain a better understanding of the actual scenario. The CHiME database has the computer technique listed (multi-source environment).

3.8 Results and Conclusions

In this part, experiments are used to compare the possible outcomes for each element of the suggested plan. The effectiveness of the suggested system will also be carefully evaluated in a variety of test scenarios. Let's begin by considering ways to improve the performance of separation and processing during detection. Furthermore, in Experimentation 1, we evaluated the proposed separation strategy using typical indicators (understandability and separability). Experiment 2 tested the suggested system under more challenging real-world conditions as opposed to the existing data set. The suggested system retains its original characteristics as well. an individual signal We confirm certain information (e.g. sentiment data).

3.8.1 CHiME Dataset:

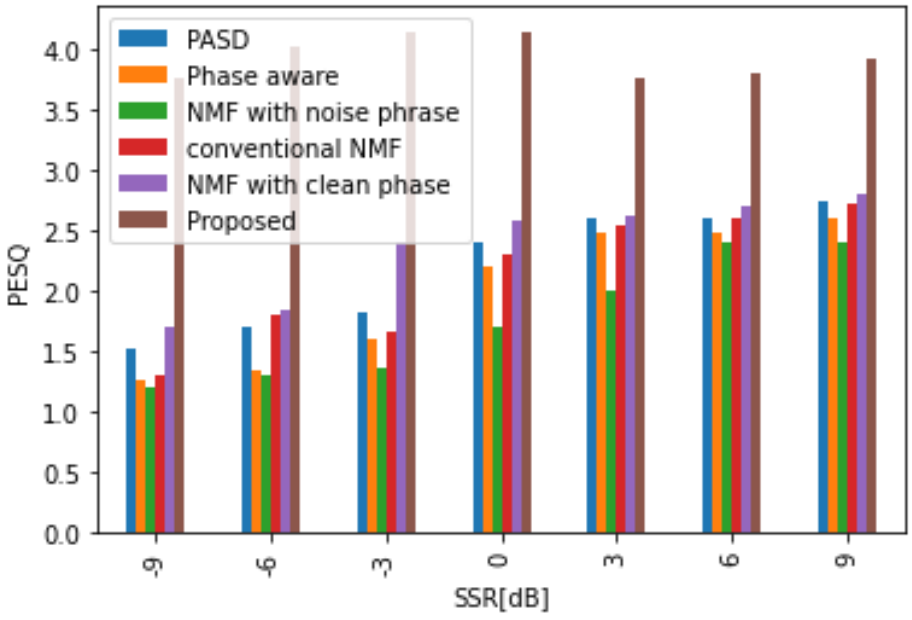


Figure 3.4: SSR evaluation comparison and study of quality (PESQ)

In terms of PESQ quality, the SSR benchmark test is depicted in Figure 3.4. Compared to competing approaches, our method can improve a single signal's perceived speech superiority. The graphic above shows the suggested method using the current pure phase NMF, Normal NMF, NMF with noise phase (relative to phase), and PASD. Examining proposed Figure 3.4 results in the greatest PESQ score. Our concept outperforms alternative approaches in terms of results.

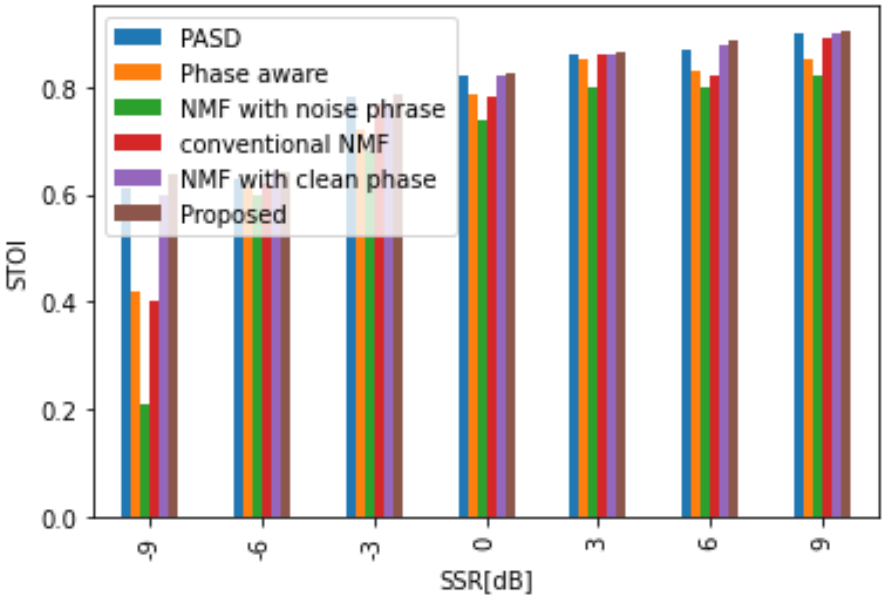


Figure 3.5: Comparison analysis of Short-Time Objective Intelligibility

The STOI benchmark test is shown in Figure 3.5, and it is used to determine speech intelligibility in loud settings. STOI and speech intelligibility have a strong association (0.79), demonstrating the usefulness of this interpreter algorithm routine. The proposed system yields superior outcomes in comparison to competing approaches. This shows that even in a loud environment, the suggested technique aims to improve speech intelligibility and phrase recognition capabilities. Among the choices are pure phase NMF, conventional NMF, noisy phase NMF, conscious phase, and PASD. The proposed Figure 3.5 will be examined to determine the highest STOI score possible. When compared to other existing solutions, our suggestion produces superior outcomes.

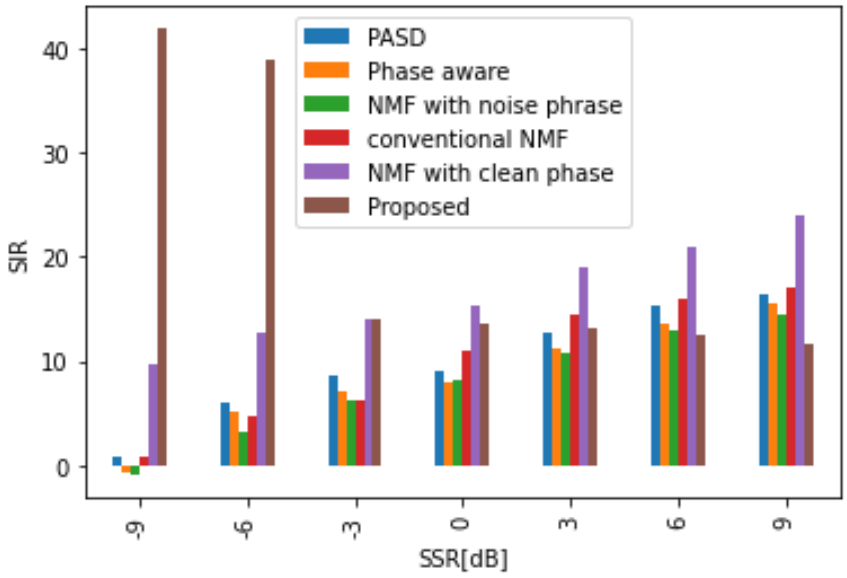


Figure 3.6: Comparison analysis of Signal to interference ratio

SIR assesses the interference rejection rate in Figure 4, which displays the results of its benchmark analysis. The suggested strategy performs better than the rival method in nearly all test circumstances. This shows that by removing interference, the suggested system delivers valuable separation outcomes. It is suggested to use a current NMF technology that includes PASD, conventional NMF, noise phase, consciousness phase, and pure phase NMF. The best SIR findings can be attained by analysing the proposed Figure 3,6. Our solution yields superior outcomes in comparison to other existing strategies.

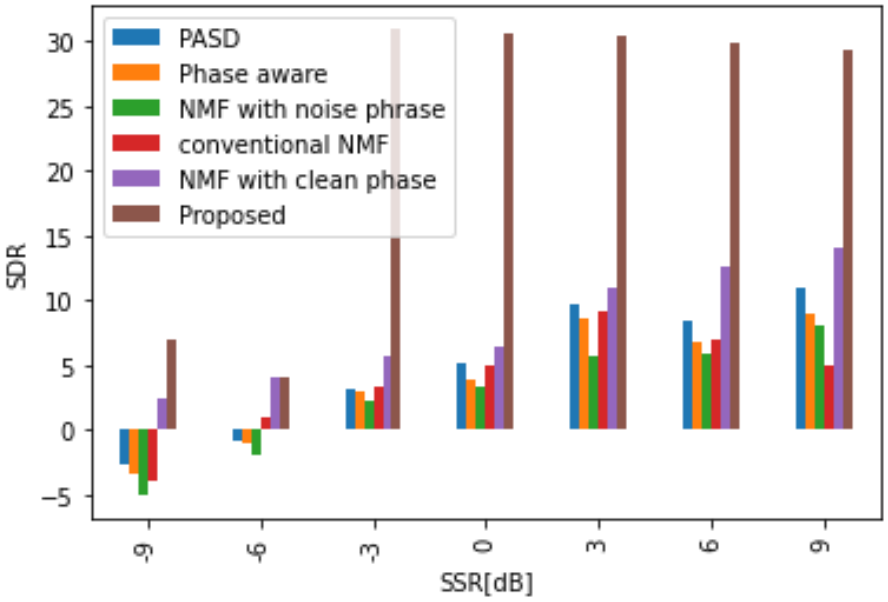


Figure 3.7: Comparison analysis of SDR

The SDR benchmark test, which has two advantages, is shown in Figure 3.7. At first, it evaluates signals separated by modified measures and in a variety of test scenarios (mixing different SDR ratios and real records). The expected results will apply to other applications. In the image above, the method for illustrating NMF using pure phase NMF, regular NMF, noisy

phase, conscious phase, and PASD is shown. Analysis of the top 3.7 results will yield the greatest SDR score. With the additional strategies we advise, better results will be attained.

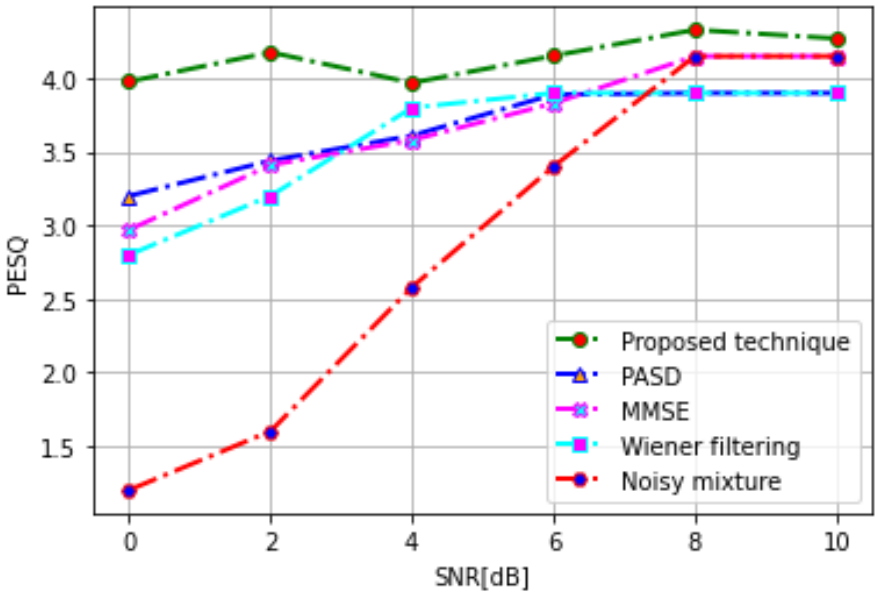


Figure 3.8: PESQ comparative analysis for the real-world situation at various SNR levels

Figure 3.8 illustrates the comparison study of PESQ results at various SNR levels in the real-world scenario of using a clean phase for reconstruction. voice indication In the image above, the method for illustrating NMF using pure phase NMF, regular NMF, noisy phase, conscious phase, and PASD is shown. By examining the provided figure 3.8, locate the SNR result with the highest SNR. Compared to other existing approaches, our technique produces better results.

3.8.2 Experiment 2: Evaluation in Terms of ASR

Here, we have calculated how well our system performs in a variety of mixed speech recognition tasks, such as those involving speakers from the same stalker (ST), speakers of the same gender (SG), and speakers of diverse genders. the mixture of genders (DG), average accuracy When compared to the baseline, the findings obtained show a noticeable improvement in performance. The proposed system performs significantly better than the fundamental system in terms of average accuracy. The evaluations of the ASR Accuracy Tables 3.1, 3.2, 3.3, and Table 3.4 yielded the following conclusions:

Table 3.1: Analysis of mixtures from the same talker in a table

Methods	**-9dB**	**-6 dB**	**-3 dB**	**0dB**	**3 dB**	**6 dB**
Baseline Signal	9	17	23	29	43	66
Advance front-end	1.5	22.0	25.5	27	46.5	68.0
SS	17	18.3	19.2	23.8	28.1	32.6
PASD	19	20.2	24.0	30.4	42.0	72.3
Proposed	82.907	82.90781	83.90781	83.90781	83.90781	83.90781

Table 3.2: Table analysis of mixtures of speakers of the same gender

Methods	-9 **dB**	-6 **dB**	-3 **dB**	0 **dB**	3 **dB**	6 **dB**
Baseline Signal	9	17	23	29	43	66
Advance front-end	9.7	14	21.4	22	46.1	66
SS	13.0	12.5	15.5	21.8	28.5	31
PASD	20.5	37.1	58.3	64.4	72.0	78.7
Proposed	90.29167	90.29167	90.29167	90.29167	90.29167	90.29167

Table 3.3: Table analysis of mixtures of speakers of the same gender

Methods	-9 **dB**	-6 **dB**	-3 **dB**	0 **dB**	3 **dB**	6 **dB**
Baseline Signal	9	17	23	29	43	66
Advance front-end	5.1	11	17	25.4	43.3	58.0
SS	14.4	18	22	27.5	36	44
PASD	37	46	62.5	70.2	75.2	80.3
Proposed	98.34219	98.34219	98.34219	98.34219	98.34219	98.34219

Table 3.4: Table analysis of Mean accuracy

Methods	-9 **dB**	-6 **dB**	-3 **dB**	0 **dB**	3 **dB**	6 **dB**
Baseline Signal	8.5	11.3	19.7	30.5	45.0	65.2

Advance front-end	7.84	11.03	20	33	45.45	65.07
SS	17	15.3	19	23	25	35.5
PASD	22	34	46	61	63	76.4
Proposed	85.87552	85.87552	85.87552	85.87552	85.87552	85.87552

Current and suggested values are analysed in Tables 3.1 through 3.4 above. In this instance, the value from the table is contrasted with the present SS, PASD, extended interface, and basic signal. Examine the proposed settings in the first four tables to attain the best ASR accuracy. Compared to other existing approaches, our technique produces better results.

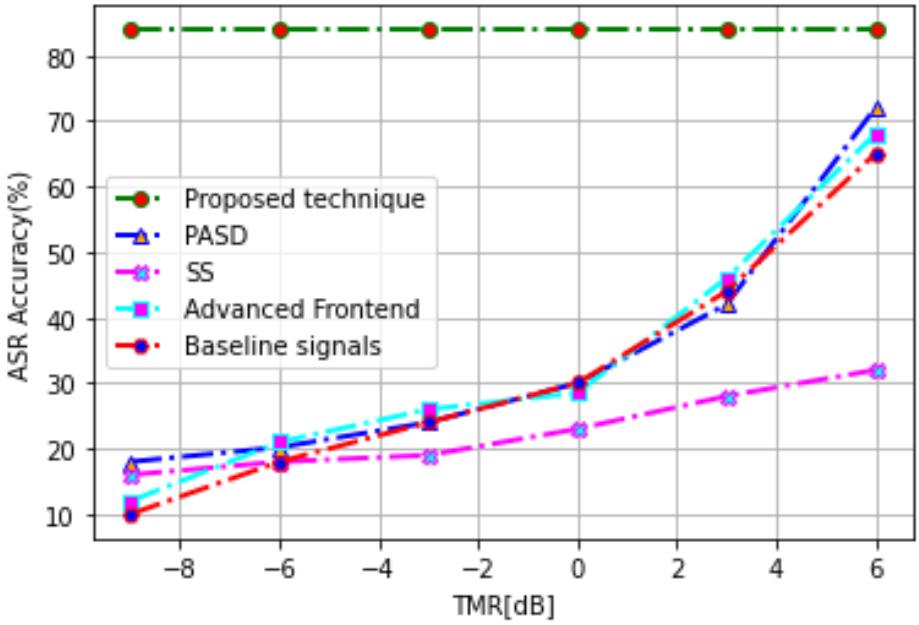


Figure 3.9: Comparative study of mixtures that belong to the same talker

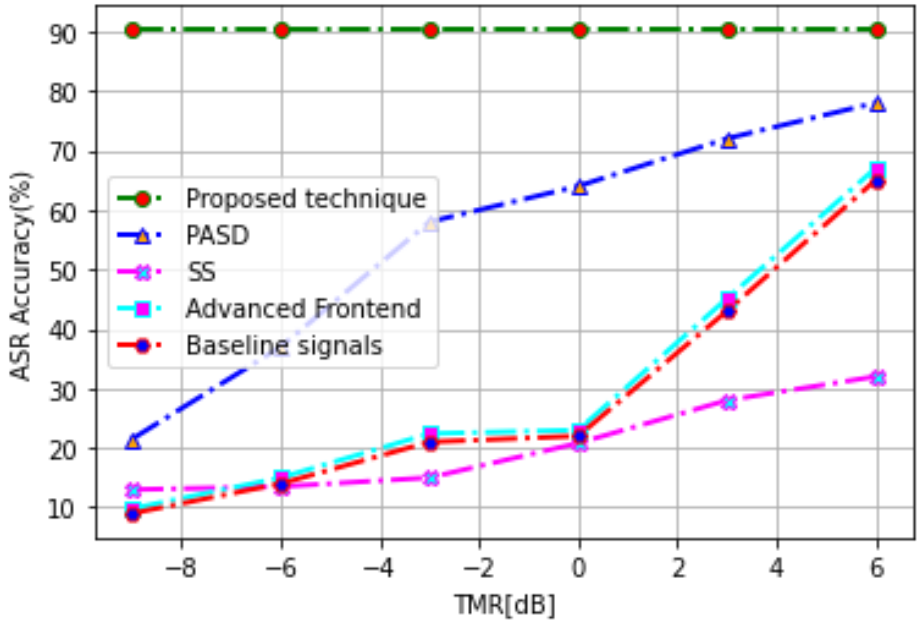


Figure 3.10: Study of comparisons between speakers of the same gender

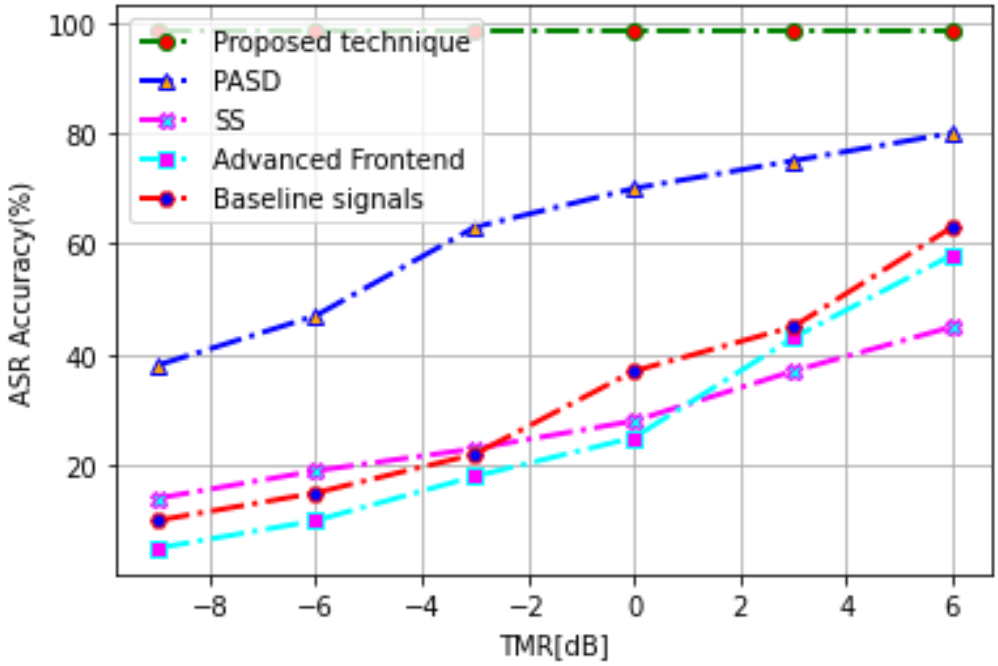


Figure 3.11: Comparative analysis of speaker combinations from different genders

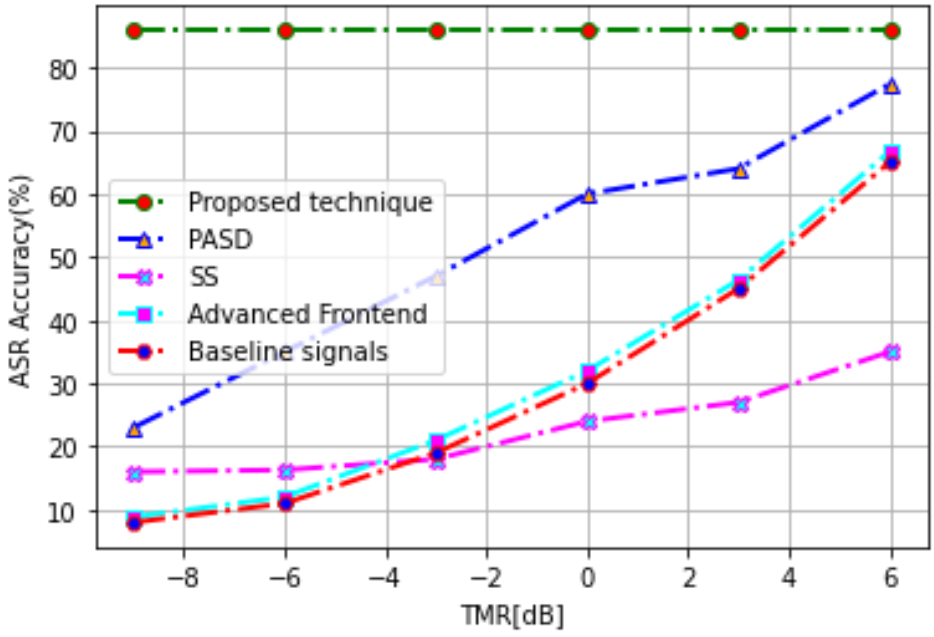


Figure 3.12: Comparison and mean accuracy analysis

The aforementioned diagram illustrates the difficulties voice recognition encounters when dealing with different mixtures, such as mixtures of speakers from the same talker (ST), mixtures of speakers of the same gender (SG), and mixtures of speakers of different genders (DG). Figures 3.9, 3.10, and 3.11 use the speech intelligibility index (SII) and the non-intrusive speech quality and intelligibility (NISQI) as markers. IBS can forecast speech comprehensibility in a variety of loud conditions and the ability to recognise sentences in adverse acoustic environments. NISQI has the highest correlation value according to subjective testing as well. the SS, front end, and PASD.

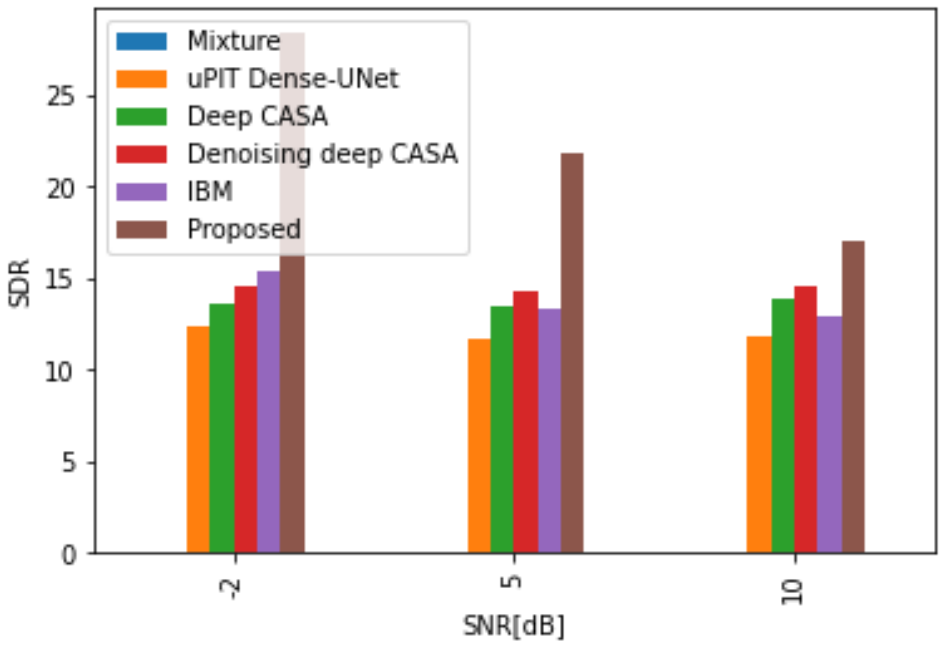


Figure 3.13: Analysis of SNR-based SDR comparisons

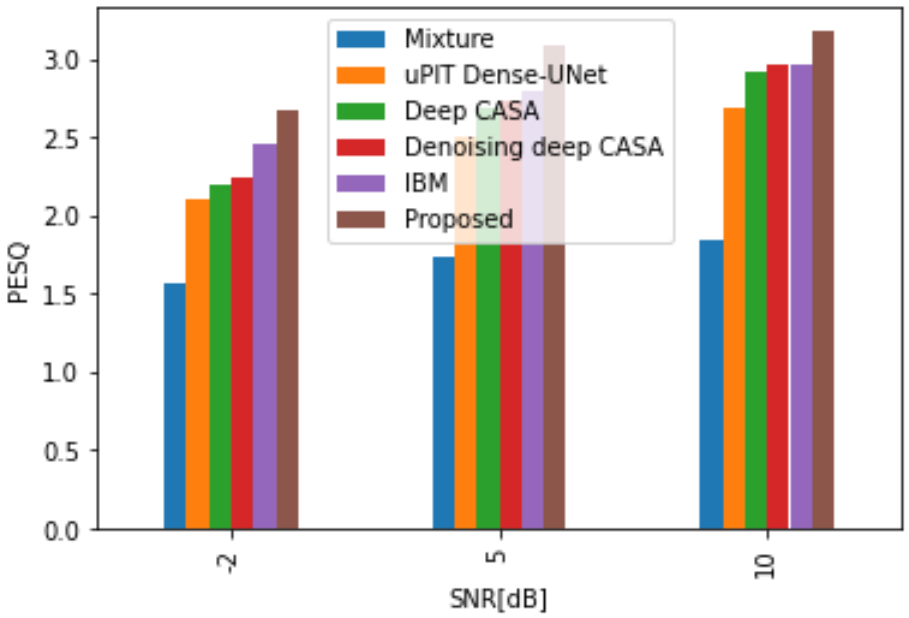


Figure 3.14: Analysis of SNR-based PESQ comparison

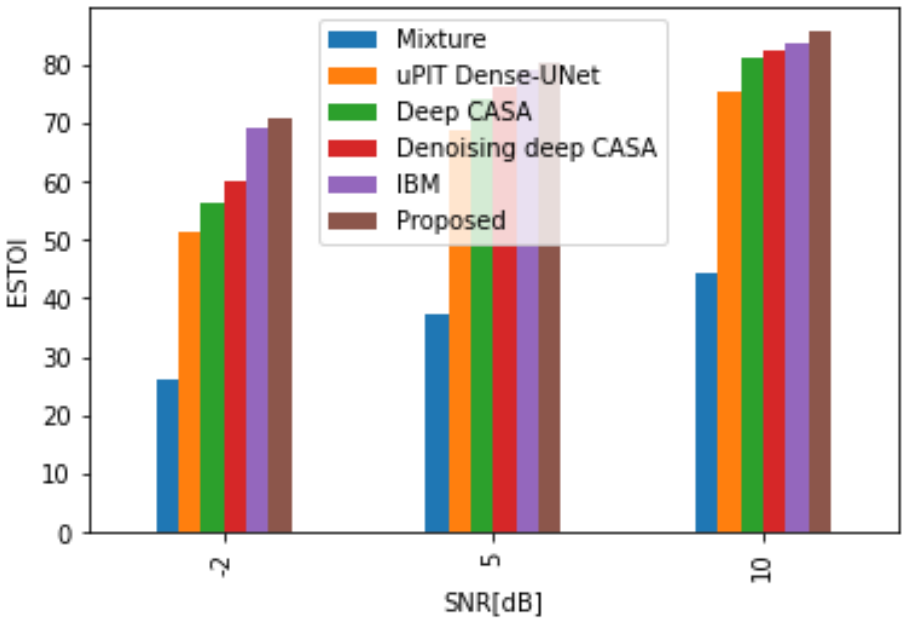


Figure 3.15: Analysis of SNR-based ESTOI comparisons

In the preceding figures 3.13 to 3.15, we contrasted our proposed method with three currently used techniques: uPIT Dense-UNet, Deep CASA, and De-noising Deep CASA, IBM. In this instance, Dense-UNet-Framework learns and employs uPIT SNR while Deep CASA is aggregated. It is a simple Deep CASA extension in a noisy environment. Remove CASA: To reduce interference from outside sources, use a lightweight version of Dense-UNet with 32-channel SDR, PESQ, and ESTOI size. Compared to other existing approaches, our technique produces better results.

3.9. Summary

In this proposed and completed part of the research work, the data consists of two or more than the clean speech signals. The TFNMF integrated with SNDNN technique has been applied and results have been obtained. It has been observed that the results (performance evaluation metrics) are improved compared with existing works as indicated in the graphs and tables. The SSR average rate is above 40 and the mean accuracy of baseline signal in 6dB is 65.2, the mean accuracy of Advanced front-end signal in 6dB is65.07, the mean accuracy of SS signal in 6dB is35.5. the mean accuracy of PASD signal in 6dB is 76.4 and the mean accuracy of the proposed signal in 6dB is85.88,

Chapter 4

Integral Fox Ride Optimization (IFRO) algorithm and Retrieval-based Deep Neural Network (RDNN) for Single Channel Source Separation

4.1 Introduction

In the noisy speech signal environment for a single channel, there is a requirement of speech signal segregation from noise. Thus, the speech signal is retrieved after getting segregated from noise. A hybrid deep neural network (HDNN) model is proposed as a unique technique for speech segregation from stationary noisy audio signal without labels. In this problem of research work, an integral fox ride optimization (IFRO) technique has been used for effective reconstruction of a variety of spectrum features which include time dynamic data, binaural and mono features. Further a hybrid retrieval-based deep neural network (RDNN) has been used for classification of speech and noise segregation.

The most instinctive form of human-machine communication is speech. Speech has become widely used in numerous applications for close-range human-machine interaction due to the dramatic recent development of speech perception (hearing and understanding) and speech generating (speaking) technologies. Reverberation, background noises, and interference speech can make speech perception (speech recognition) and speech creation (text-to-speech) systems less accurate.

4.2 Single-Channel Speech Separation

Single-channel speech separation is the process of estimating numerous output waveforms from a single input recording in which multiple speakers are speaking at the same time, each estimate having the voice of only one of the input speakers. Single channel speech separation must use only the structure of speech and must capitalize on inter-speaker differences, relying heavily on the fact that each speaker's speech is sparse in a time-frequency domain. This is in contrast to multi-channel techniques, where multiple microphones capture the speech and provide access to directional information. In other words, it is unlikely that numerous speakers will contribute a large amount of energy to a segmented signal if a combination of speakers is segmented spectrally, for instance using a straightforward Short-

Time Fourier Transform (STFT) with suitable settings. The latent speech signals are now simpler to recognize during training and inference, in addition to making it easier to split the signal into a spectrum representation [100]. Source separation strategies were frequently based on either known features of the speech signals or inspired by the human auditory perception system's capacity to follow sources in overlapped speech prior to the development of DNN-based methods propelled by massive volumes of labelled data.

These traditional techniques include Computational Auditory Scene Analysis (CASA), Factorial Hidden Markov Models (HMMs), Independent Component Analysis (ICA), and Non-negative Matrix Factorization (NMF). These techniques are frequently based on statistical features of the signals and rely on signal processing to distinguish between the sources. Comparing this class of speech separation approaches to other separation tasks like reducing speech noise, the major problem is that speech signals from two different speakers can have extremely similar statistical features. The algorithms do take advantage of the structure and continuity restrictions of speech in time and frequency, which results in some success in speech separation, but their performance is considerably outperformed by the more recent deep learning techniques.

The bulk of DNN speech separation approaches rely on a spectral masking strategy, however some cutting-edge methods estimate the speech source waveforms directly. In order to use these techniques, the mixed waveform must first be projected using an analytical transform into a two-dimensional spectral domain with resolution in both time and frequency. The Short-Time Fourier Transform (STFT), which was employed in earlier approaches, has now been replaced by learnt transforms.

Then, using this mixture spectral representation, a neural network creates a mask for each speaker with values ranging from 0 to 1. An approximation of the source spectra of individual speakers is produced by individually multiplying each of these masks by the mixture representation in order to mask out the interfering sources.

4.3 Proposed method

The first issue, over-smoothing, is addressed, and estimated signals are used to expand the training data set. Second, DNN generates prior knowledge to address the problem of incomplete separation and mitigate speech distortion. To overcome all current issues, we suggest employing an efficient optimal reconstruction-based speech separation (ERSS) to overcome those problems using a hybrid deep learning technique. First, we propose an integral fox ride optimization (IFRO) algorithm for spectral structure reconstruction with the help of

multiple spectrum features: time dynamic information, binaural and mono features. Second, we introduce a hybrid retrieval-based deep neural network (RDNN) to reconstruct the spectrograms size of speech and noise directly. The input signals are sent to Short Term Fourier Transform (STFT). STFT converts a clean input signal into spectrograms then uses a feature extraction technique called IFRO to extract features from spectrograms. After extracting the features, using the RDNN classification algorithm, the classified features are converted to softmax. ISTFT then applies to softmax and correctly separates speech signals.

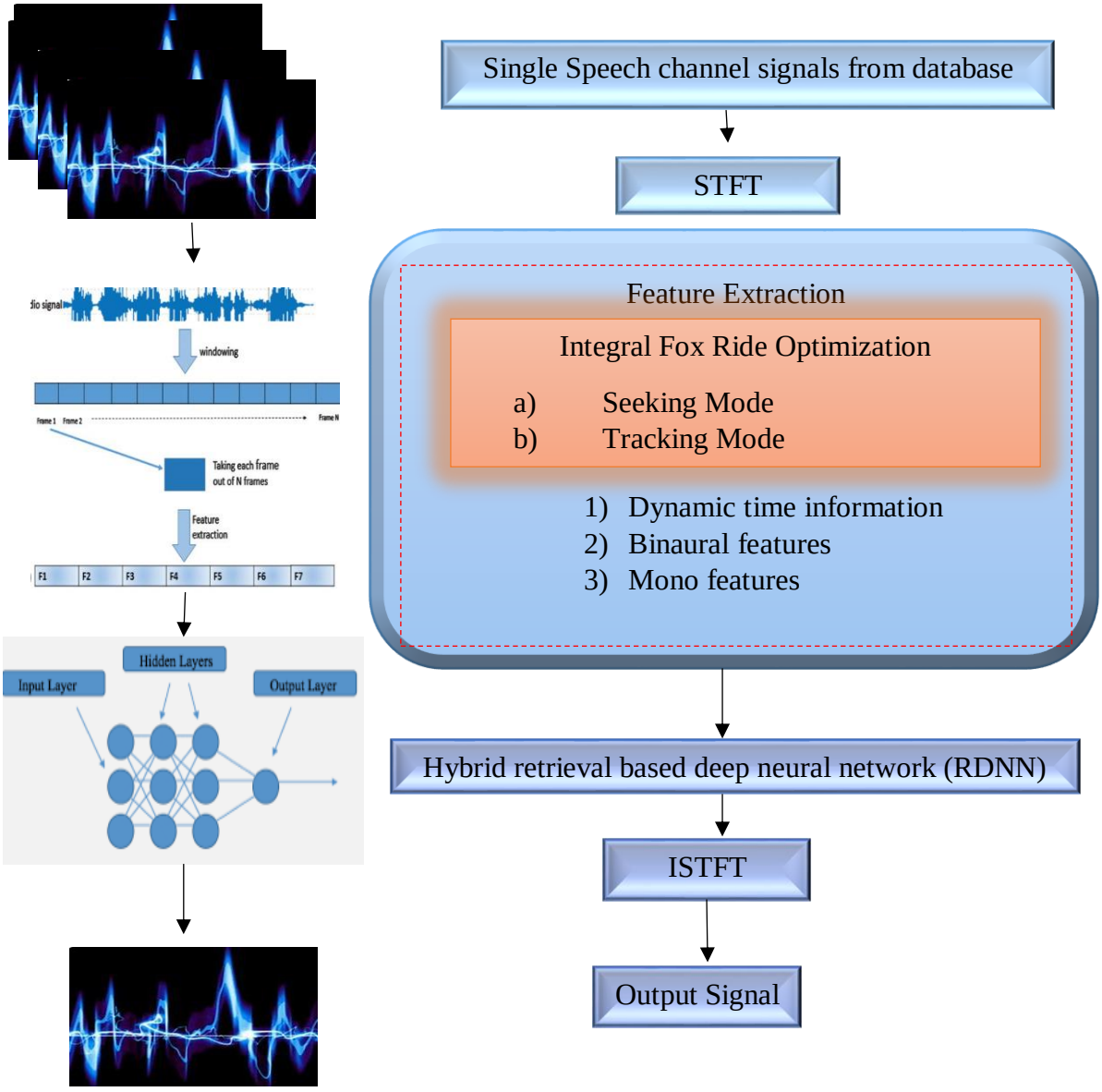


Figure 4.1: Block diagram of Integral fox ride optimization based RDNN system

We propose an efficient optimal reconstruction-based speech separation (ERSS) using a hybrid deep learning technique. The objective of feature extraction is to improve the quality of the

training data set extracted from each of the speech signals from low-level texture features using integral fox ride optimization. The output from feature extraction was given to segmentation and in a cascaded process to provide a textured pattern. Finally, using the RDNN classifier, we found the noise error from speech and removed it using the proposed technique.

When we use NMF to capture the structure patterns of speech separation targets, such as ideal masks or magnitude spectrograms of interests, We investigate a discriminative training objective with sparsity constraints, which improves the separation model's ability to suppress noise and preserve speech.

NMF is used to understand the important spectrum of speech and sound, whereas DNN is used to assess the essential spectrum's function. The NMF hypothesis and functional assessment are combined with DNN to comprehensively reproduce clear sound and sound within the compound. The combined strains of DNN and NMF are improving the performance of the voice department. We suggest a different optimization range with interval control to suppress excessive noise. This reduces the residue of isolated speech and noise and dramatically improves GSIR performance. Models can stop high interactions and outperform comparative models with very low-cost hand tools and defects. Production models can use spectral structures based on speech and sound, while in-depth study models study complex linear graphs of distinguishing objectives through silent and supervised learning.

The latest approach since optimizes training formal speech segmentation, in which different modes of speech, speaker, and background sound are studied from training data. Several supervised separation systems have been proposed. The in-depth learning methods used for supervised speech separation increased the rate of progress and increased the separation efficiency [101]. Also, reliable assessment of time-frequency masks from the conversation is challenging, especially when there is room echo in the mix.

We propose an efficient optimal reconstruction-based speech separation (ERSS) to overcome those problems using a hybrid deep learning technique.

- First, we propose an integral fox ride optimization (IFRO) algorithm for spectral structure reconstruction with the help of multiple spectra features: time dynamic information, binaural and mono features.

- Second, we introduce the Deep Neural Network (RDNN) based on a hybrid search to directly reproduce the speech and voice level spectrogram. RDNN can instantly improve the partitioning range and minimize accumulated errors.

• Finally, we implement the proposed design in the MATLAB tool, and the performance of the proposed ERSS is compared with the existing state-of-art techniques.

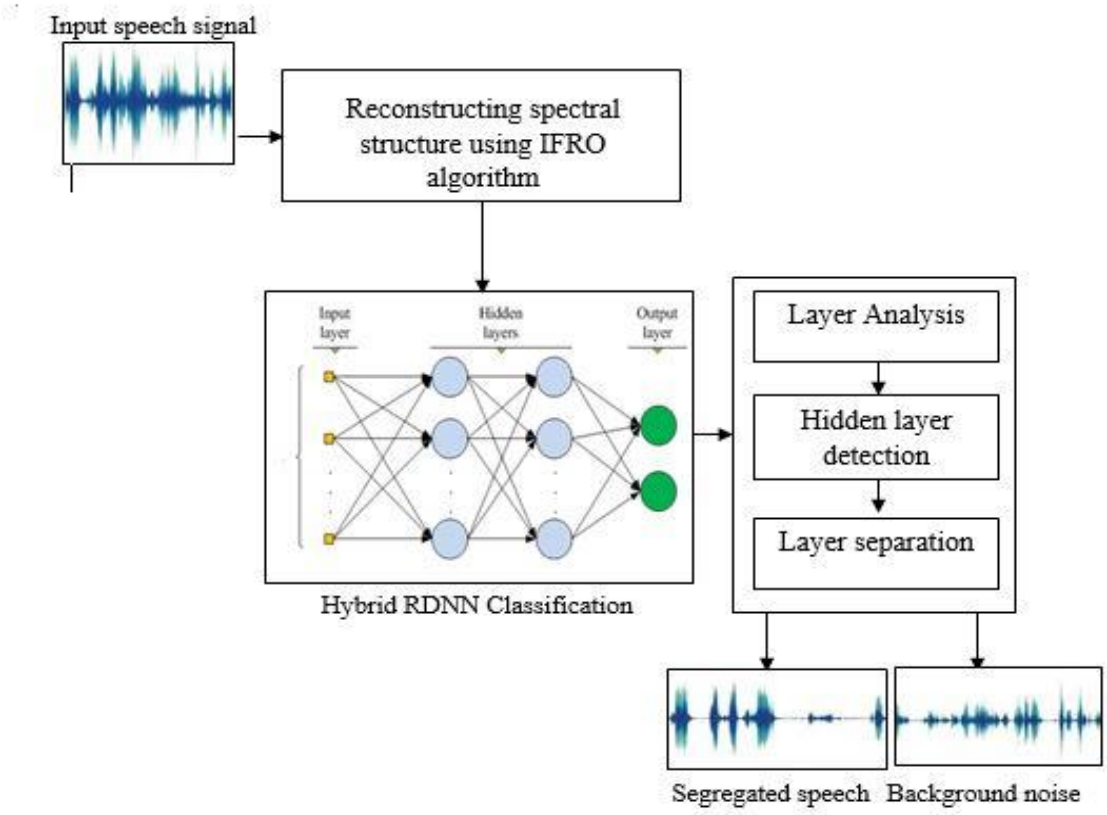


Figure. 4.2: Illustration of proposed ERSS using a hybrid deep learning technique

Figure. 4.2 shows a specific hybrid model structure separating background noise and conversation using Fox Riding optimization and Search Deep Neural Networks. As shown in Fig. III, a multi-layered deep neural network input speech signal extract with continuous functions such as non-linear activation, regulation, and hidden layer to extract advanced features of the speech signal. After the fully connected hidden layer's input layer, a multi-dimensional speech signal was extracted from the connected first layer. Finally, the classifier separates the background noise from the speech signal.

4.4 Proposed optimization and deep learning technique

4.4.1 IFRO optimization

The FRO system has two modes, i.e., the searching for away and the going with way. In checking for mode, Fox search for in their neighbourhood, which has a spot close to looking. Fox updates their condition in the following method by following the general faultless worth, an area with the available seek after. FRO has a solid combination, both thinking about worldwide enhancement and nearby streamlining, which is for the most part connected to work improvement and accomplished significant impact.

a) Seeking mode

The looking for method portrays the resting aptitude of a fox. A fox moves to various positions in the inquiry space, looking for a way yet stay alert. It very well may be translated as nearby look for the arrangements. The accompanying documentation is utilized in this model

- The searching Ratio of chosen Dimension (SRD) signifies the distinction among new and old components of fox chosen for change.

- Searching Memory Pool (SMP): This parameter portrays the number of duplicates of a fox to be reproduced.

- Dimension Counts Change (DCC): It speaks to the number of measurements a fox position experienced for transformation. The means of seeking a method of FRO calculation are given as pursues.

If SPC=I, Generate T (=Searching Memory Pool) copies of as indicated by DCC, request the change administrator to the T duplicates. Arbitrarily short or in addition to Searching Ratio of selected Dimension percent the present qualities, supplant the old attributes [102]. Assess the wellness of the changed duplicates. Use condition (1) to compute the choosing likelihood of every competitor and pick the point with most elevated choosing l likelihood to supplant. If the objective of the wellness capacity is to locate the base arrangement, Le , otherwise

$$P_i = \frac{\left|FS_i - FS_b\right|}{FS_{\max} - FS_{\min}} \qquad (4.1)$$

b) Tracking Mode

Tracking Mode is the second method of calculation. In this mode, felines want to follow targets and nourishments. This mode mirrors the chasing ability of felines. When a feline the prey, the position and speed of the feline are refreshed. This way, an enormous contrast happens between old and new places.

Representation of the best position of a fox is that the fox's position and velocity are calculated using (1) & (2) equation.

$$\alpha(t) = \alpha_{\min} + (\alpha_{\max} - \alpha_{\min}) \sin(\frac{\lambda t}{t \max}) \tag{4.2}$$

Where, new described the refreshed velocity of dimension, indicates fox dimension and w indicates a factor weight from the value of 0 to 1, shows the past velocity of the fox term, c represents user finite number,

Where, $V_{inew}^{\ d}$ new described the refreshed velocity of d^{th} dimension, i^{th} indicates fox dimension and w indicates a factor weight from the value of 0 to 1, $V_{ibest}^{\ d}$ shows the past velocity of the fox i^{th} term, c represents user finite number,

$$P_{jnew}^{\ d} = P^{d}_{\ j} + V_j^d \tag{4.3}$$

Where, now indicates the position update of fox in dimension, shows the present state position with fox and size then denoted the fox velocity of the term. To investigate additionally encouraging arrangement and develop a ratio of convergence, while the fox best position is utilized to control the places of fox in the following mode. Subsequently, another changed quest condition is implemented for the following method of FRO calculation, which incorporates the worldwide best [34]- [36]

$$P_{inew}^{\ d+1} = (1-\alpha) * P_j^{\ d} + \alpha * N_g + V^d_{\ j} \tag{4.4}$$

The FRO calculation utilizes a speed vector, and past fox locations were refreshed in tracing mode. The restored fox location is just affected by vector velocity. Thus, another speed refreshed condition is presented to develop an assorted variety of FRO calculations, particularly within finding mode.

$$V_{jnew}^{\ d+1} = V_j^{\ d} + \alpha\left(N_g - P_j^{\ d}\right) + \beta * \varepsilon \tag{4.5}$$

where, ε is an irregular vector consistently conveyed from [0 to 1]; α and β are quickening parameters used to sift through the state of a feline toward close to better positions, and Pg provides the general position for the best situation of a feline. To concordance between the appraisal and misuse structures, both vitalizing parameters β and α go about as parameters controlling.

$$\beta(t) = \beta_{\max} - \{\frac{\beta_{\max} - \beta_{\min}}{t_{\max}}\} * t \tag{4.6}$$

In (4.6), presenting the lower and upper limits, t indicates the most extreme no. of cycles, and t denotes the present emphasis value. Subsequently, α (t) is a stage work whose worth ranges among lower and upper limits. The bigger estimation of α bolsters investigation, whereas little qualities bolster abuse. The point of α (t) term is investigated and controlled by the procedure of fox in hunt space.

$$\alpha(t) = \alpha_{\min} + (\alpha_{\max} - \alpha_{\min}) \sin(\frac{\lambda t}{t \max}) \tag{4.7}$$

In (4.7), the mean the base and most extreme estimations of first and last cycles respectively represent the greatest no. of emphases and t described the present value in iteration. The explanation for the consolidation of the parameter is to impact the worldwide investigation capacity of the proposed calculation; a massive estimation of the parameter reinforces the worldwide best position of feline and watches out for the arrangement refinement. The pseudocode of FRO is shown in Algorithm 4.1.

The algorithm 4.I Pseudocode of integral fox ride optimization

Input: A speech signal with background noise

Output: Separate noise and speech

1	Initialize the various parts of proposed count like sum α, neighbourhood structure, fox (N), β, C, SRD, SMP, and C are variably put N addresses a position in numbers in sporadic search space.
2	Generate every cat from the D-dimensional space of search speed and position.
3	Compute the fox wellbeing limit and save an estimation value, which is the best position.
4	While (i < mi)
5	To evaluate the no. of Flag, distributed randomly seeking along tracing mode towards the fox.

6 If (Flag==1); Seeking mode of fox position

7 To apply the seeking mode to every fox.

7.1: Generate every fox j copy.

Maintain the fox best position after the contrast value of fitness function toward memory.

End for

8 Else, tracing the mode position of the fox

To apply the tracing mode to every fox

Find the fox best position after update the fox position

rand ≤ iterfittness

Update global best position and fox position.

End if

i = i++

Obtain the concluding solution

4.4. 2 Hybrid retrieval based deep neural network

This segment describes the systematic description of retinal-based deep neural network (RDNN) and the creative learning process for dynamic DNN generation. Finally, the cumulative criteria are given.

Systematic Description: The repaired deep neural network with concealed layers m can be reported using 2 constraints (Λ, Φ); Λ as shown in Figure III. Layers are vectors that give the figure of neurons per layer 1; = Λ = (n0, n1, ..., nl , ..., nm, nm+1). The input layer is (l = 0), the output layer is (l = m+1), and the hidden layer is (l to m). = W1, W2,..., Wl,..., Wm + 1= Φ is the weight connections vector. Each of the W1 vector's components is a heavier link matrix.

$$W^{l} = \begin{pmatrix} w_{11}^{l} & w_{1j}^{l} & w_{1nl-1}^{l} \\ w_{il}^{l} & w_{ij}^{l} & w_{inl-1}^{l} \\ w_{nl1}^{l} & w_{nlj}^{l} & w_{nl\,nl-1}^{l} \end{pmatrix} \tag{4.8}$$

Where is the weight correlation between the ith neuron of the layer l and the jth neuron of the layer l − 1. Consider only RDNN with one neuron at the output layer + + 1 = 1, where d is the input vector size (input layer size n0 = d) to simplify the technique described in the following article.

Specific evolutionary simulation RDNN is a neural network structure that regularly develops through training. The size of layer l is denoted by the vector n t l in phase t, while the vector Λ t indicates the size of layer t. The fundamental structure of the neural network at the start of the creation phase is: Λ 0 = n 0 0, n0 1, n0 2 = (d, 1, 1). Figure 1 shows how to create the first hidden layer (HL 1) from scratch. The construction procedure is separated into three sections at each step.

- In the primary stage, new neurons are included. This novel neuron is totally integrated with all the preceding and subsequent layers.
- We are starting a new burden. All other weights retain their preceding values.
- At last, HL (1) only trains the weight of the concealed layer and constantly updates it using the online backdoor algorithm.

The second concealed layer is built in the same way as the first concealed layer HL (L).

Full description: The evolutionary architectural algorithms and how to achieve the integration criteria are described in the following section. At every training stage of layer l, the training process is completed by reducing the M1 frequency of the W1 to continuous online backs to update the weight, with the square error function defined by X at each repetition (from = 1 to M).

$$E_K = \frac{1}{2}(o_k - d_k)^2 \tag{4.9}$$

Where ok represents the neural network output for the K format, represents the output required for the K format, k represents the code above the input-output pair (k = 1 to N), and N represents the number of samples in the training set. To update the W1 weight, calculate the (1) slope utilized in the random online back spreading method. By adding to it, the weight is updated:

$$\Delta w_{ij}^{l} = -\eta \frac{\partial E_k}{\partial w_{ij}^{l}} \tag{4.10}$$

Where is the development rate, (GSEiter) delivers the total square error of the N training pairings at the end of each iteration:

$$GSE_{iter} = \sum_{k=1}^{N} E_k \tag{4.11}$$

At the end of the step-by-step training process, the average square error (MSEt) returns.

$$MSE^{t} = \frac{1}{M}\sum_{iter=1}^{M} GSE_{iter} \tag{4.12}$$

For each of the four scenarios in the building process, keep the following information in mind:

- It's not reached. It is objective and not completed with hidden layer l . A hidden layer was added with a new neuron
- It's not called its objective and full l hidden layer. A hidden layer was added with a new neuron.
- Reached its objective. Successfully built a DNN. End of the building process.
- RDNN reached its peak regardless of the end goal. RDNN is not built successfully at the end of the construction process.

Where θ is the gateway used to define RDNN, is the number of neurons in T according to the currently concealed layer l, and is the number of neurons in T. The maximum RDNN recognition is called Max. Maximum hidden layer; The complete RDNN has approved the maximum number of hidden layers, Max. Authorized layers for the entire RDNN.

Algorithm 4.2 Initialization process of RDNN

```
1    Part 1: Initialize RDNN process
2    t = 0,
3    l = 0,
4    Max1, max. no. of Hidden Layers
5    Max = random, max. no. of neurons per layer
6    // DNN initialization7
8    // initialize random no. of weights:
10   //end the process
```

In this way, the training course of RDNN is repeated at each stage according to the calculated MSEt. Max and Max1 are used to control the RDNN level. To avoid the arbitrary size limit of hidden layers, I utilized a random limit:

$$Max_{n} = random(\alpha_{1},\ \alpha_{2}) \tag{4.13}$$

Where $(\alpha_{1},\ \alpha_{2})$ are correspondingly the lower limit and higher limit? The functional capabilities of DNN are given in Algorithms 2-4, along with the boot process, hidden layer 1, and fine-tuning, respectively, for updating the weighted link in the last layer.

Algorithm 4.3 Building hidden layer-l in RDNN

```
Repeat
for iter = 1 to M
for k = 1 to N
calculate
calculate
 // Update the weights Wl
end k
calculate
end iter
calculate
end
// RDNN successfully Built
 // hidden layer l added to new neuron
if
// added hidden layer
if
ends
 // Not built the RDNN
t  =  t + 1
end
```

4.4.3. Comparative Analysis of Previous Speech Separation and Enhancement Work

Our evaluation would not be complete without comparing our results to previous work in an efficient optimal reconstruction-based speech separation (ERSS) using a hybrid deep learning technique. It contains these comparisons on different efficient optimal reconstruction-based speech separation (ERSS) using mixed deep learning datasets, integral fox ride optimization (IFRO) algorithm, RDNN and MATLAB tool, using the evaluation protocols and metrics described in the respective papers. On our project page, you sometimes get qualitative results from these comparisons. It's worth noting that these previous methods necessitate training a separate model for each speaker in their dataset (speaker-specific), whereas we evaluate their data using a model built on our general RDNN dataset. Despite never having

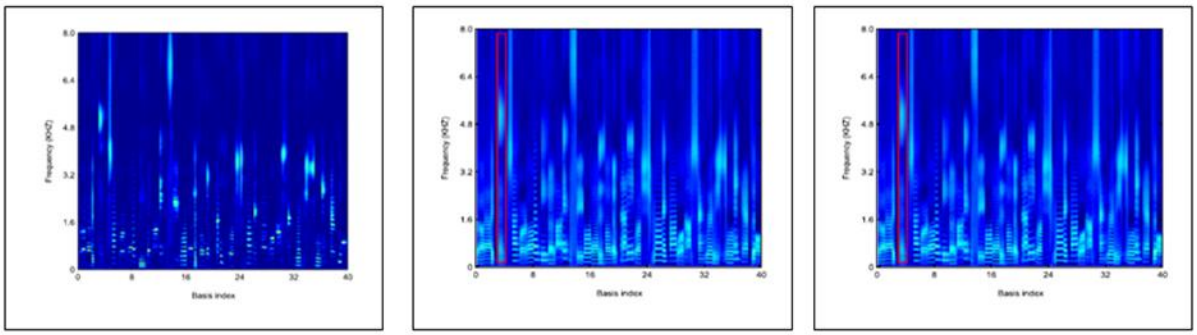
heard these specific speakers before, our results are substantially better than those reported in the original articles, demonstrating our model's great generalization capacity. We propose an efficient optimal reconstruction-based speech separation (ERSS) using a hybrid deep learning technique to overcome those problems.

4.5 Results and Conclusions

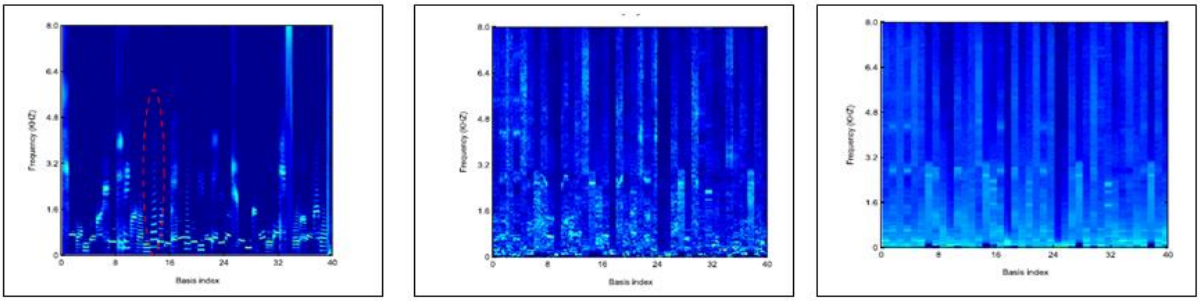
Tests were performed to evaluate the performance of the anti-supervisory control source or filter model for speech separation. Benchmarks include Semi-supervised source or filter models with variations in control usage (with or without controls, better control over the source or filter sync control, control adaptation generated for sound isolation).

4.5.1 Dataset Description

For evaluation, TIMIT Corpus and Noisex-92 Corpus are used as voice and audio data, respectively. TIMIT contains 10 sentences spoken by 630 speakers of 8 different dialect regions of the USA. The NOISEX-92 contains 15 general types of sounds in a typical environment, each about 4 minutes long. The NOISEX-92 has noise such as factory noise, F-16 noise babble noise, etc. While mixing speech and noise, we randomly cut each NOISEX-92 noise utterance into unique portions based on the time length of speech utterances to ensure that the various components of each noise utterance are mixed with the clean speech utterances. These sounds are mainly related to different everyday sounds, and they are also non-permanent. Nine types for training DNMF, SNMF, CNMF for speech, 2000 words for speech-based sound, and 2,000 words for sound training. W1 and W2 were trained with 2000 words and phonetic pairs. Figures 4.3 (a-f) show speech basis spectra and noise basis spectra.



a) b) c)



d) e) f)

Figure.4.3: (a) Typical NMF, (b) Sparse NMF, (c) discriminative NMF (trained using the TIMIT training database) found a collection of speech basis spectra. (d) TNMF, (e) SNMF, (f) DNMF (trained with 9 noises from the NOISEX-92 dataset) identified a collection of noise basis spectra.

4.5.2 Simulation setup

This article provides a comprehensive summary of RDNN supported supervisory speech separation. We reviewed the key components of the supervisory department: describing learning machines, educational goals, vocal functions, representative methods, and reviewed several related studies. DNN-based segregation and segregation issues were created as a supervised study, which has dramatically elevated modern art to some linguistic tasks, including moral language development, language disabilities, speaker segregation, and continuous sound partition. This speedy improvement will lead to a rigorous combination of domain awareness and data-based frameworks and the development of in-depth knowledge. Beneath, we converse some of the ideological issues related to this perspective.

It is important to define appropriate training goals for learning and generalization in the supervised conversation category. There are two categories of educational goals: mask-based goals and mapping-based goals. Cognitive-based objects define the time-frequency relationship for clean speech background interaction, while mapping-based objects relate to pure speech spectrum representation. This section compares the RDNN methodology to four popular NMF models: Typical NMF, Sparse NMF, Discriminative NMF, and Convolutive NMF.

4.5.3 Performance Metrics

As assessment estimations, we receive SIR, SAR, SDR, SNR, PESQ worth [-0.5, 4.5] and a brief timeframe target clearness measure (STOI [0,1]). SIR, SAR, and SDR are free to use and can be enrolled by the Blind source separation (BSS) Evaluation tool section to test degrees of basis to impedance, antiquities, and mutilation. The SNR and PESQ scores independently assess the degrees of the source to uncomfortable influence and target talk worth, whereas the Short time objective intelligibility (STOI) assesses target speech coherence.

Furthermore, we compare the SIR, SDR, SNR, PESQ, and STOI enhancements in terms of the blend talk, as follows:

$$GSIR(\hat{s}, s, x) = SIR(\hat{s}, s) - SIR(x, s), \quad (4.14)$$

$$GSDR(\hat{s}, s, x) = SDR(\hat{s}, s) - SDR(x, s), \quad (4.15)$$

$$GPESQ(\hat{s}, s, x) = PESQ(\hat{s}, s) - PESQ(x, s), \quad (4.16)$$

$$GSNR(\hat{s}, s, x) = SNR(\hat{s}, s) - SNR(x, s), \quad (4.17)$$

$$GSTOI(\hat{s}, s, x) = STOI(\hat{s}, s) - STOI(x, s), \quad (4.18)$$

Where GSIR, GSDR, GSNR, GPESQ, and GSTOI denote the gains of SIR, SDR, SNR, PESQ, and STOI, respectively. Here's' is the pure speech, x is the mixture signals, and ˆs is the divided speech. It is a method of weighing all grade measurements and test clips along their length, the higher principle indicating better performance. Furthermore, simultaneous speech and sound level spectrum prediction improves separation efficiency. On the one hand, sound and noise levels can cover a large part of the spectrum and separate sounds. Preliminary studies, on the other hand, show that the Weiner type filtering strategy can increase the overall performance of RDNN much further. Compared to the earlier mask approximate RDNN, the approximate spectral target provides several advantages.

Table 4.I Various metrics using existing and suggested techniques

Models	gSDR	gSAR	gSIR	gPESQ	gSTOI
Proposed ERSS	**10.90**	**10.80**	**15.30**	**0.58**	**0.08**
Joint-DNN-DNMF	9.90	10.40	14.60	0.54	0.07
Joint-DNN-CNMF	10.0	10.40	14.80	0.57	0.07
Joint-DNN-TNMF	10.1	10.50	15.00	0.57	0.07
Joint-DNN-SNMF	9.60	10.40	13.40	0.50	0.07
DNN-SPE-NOI-5	9.60	10.70	13.30	0.50	0.07
DNN-SPE-NOI-1	9.50	10.50	13.00	0.47	0.07
DNN-SPE-1	8.10	8.40	11.70	0.40	0.06
DNN-SPE-5	8.60	9.20	12.30	0.45	0.07
DNN-PSA-1	9.60	10.10	14.80	0.42	0.05

DNN-PSA-5	9.60	10.20	14.50	0.45	0.05
DNN-IRM-1	8.50	10.60	10.90	0.45	0.06
DNN-IRM-5	8.50	10.80	11.40	0.44	0.06

Table. 4.1 show the different models like TNMF, SNMF, DNMF, CNMF, and proposed ERSS using four execution estimations: gSAR, gSDR, gSIR, and gPESQ gSTOI. This phenomenally owes to the joint undertakings of RDNN and IFRO. As indicated by one point of view, RDNN can misuse spectra-standard structures of talk and change by taking in premise spectra from tremendous unadulterated talk and blast. On the other hand, RDNN has strong demonstrating limits in taking in the non-linear organizing from the obligation to target. The planned combinatorial game-plan centers on the qualities of the pair RDNN and IFRO for the talk group.

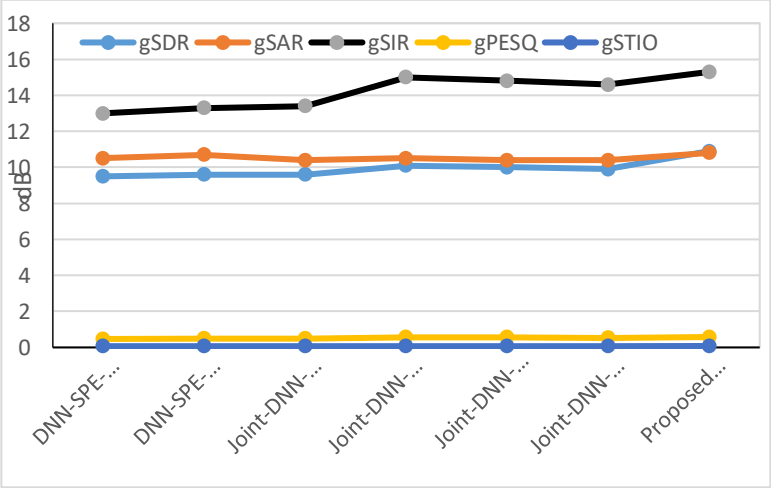


Figure. 4.4: Graphical representation of Speech Separation Performances of Various metrics using existing and suggested technique

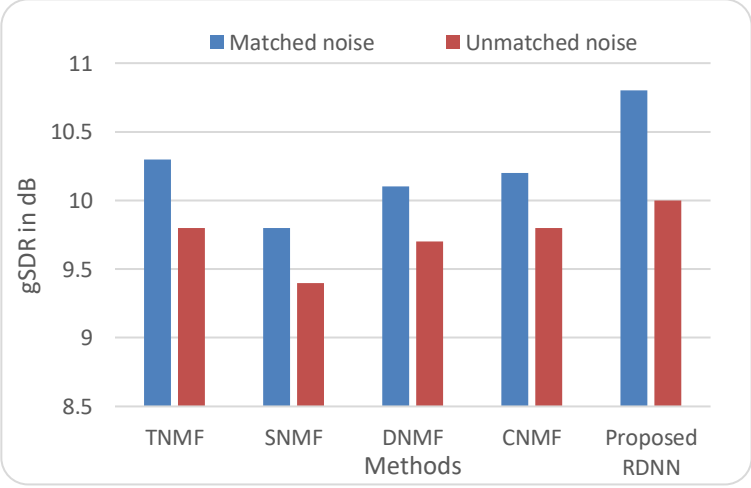


Figure. 4.5: Graphical representation of gSDR matched and unmatched noise

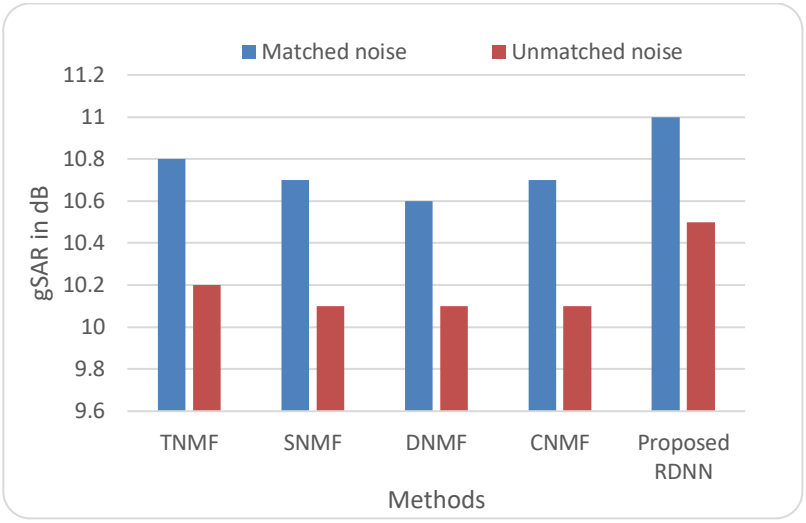


Figure. 4.6: Graphical representation of SAR matched and unmatched noise.

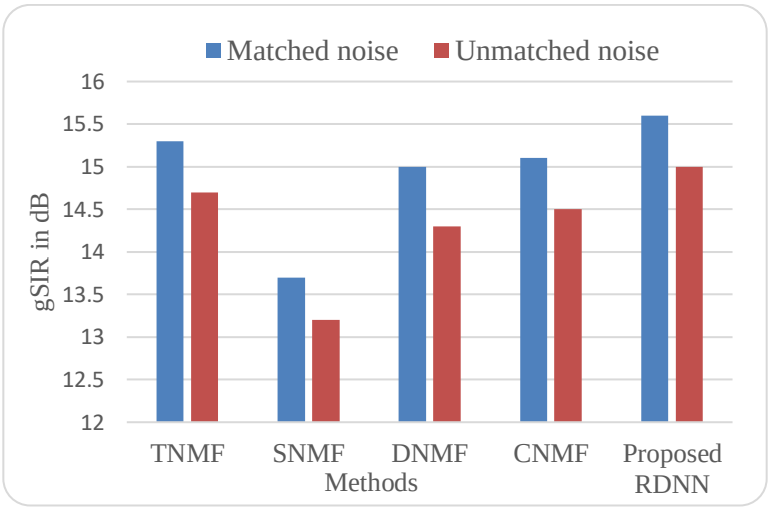


Figure. 4.7: Graphical representation of gSIR, matched and unmatched noise

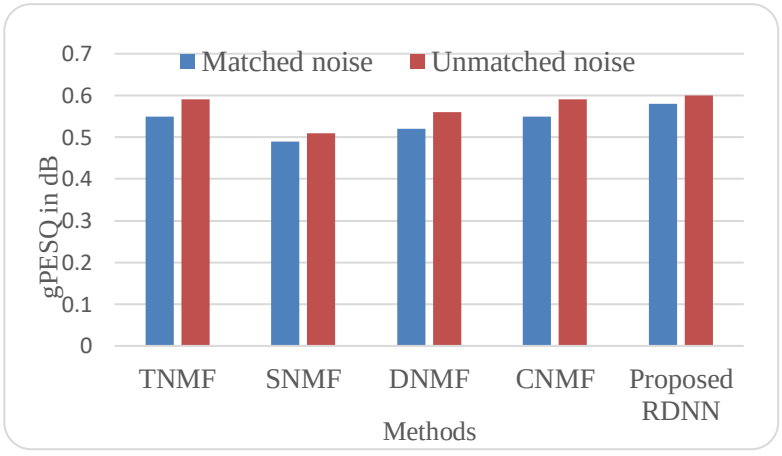


Figure. 4.8: Graphical representation of gPESQ matched and unmatched noise

Figure. 4.5 to Figure.4.8 reports the cultivated presentations by Joint-DNN-TNMF, SNMF, DNMF, CNMF with proposed RDNN for various sparsity models. From one viewpoint,

RDNN can abuse spectra-common talk and clatter structures by taking in premise spectra from enormous unadulterated talk and uproar. Of course, RDNN has strong showing limits in taking in the non-linear arranging from commitment to objective. The projected combinatorial arrangement considers the mutually RDNN and NMF for the talk segment. Although Kang-DNN-NMF also abuses the characteristics of the pair RDNN and IFRO for talk division, the IFRO indication and the RDNN measure of the authorizations are acted in an alternate or channel way. This will incite a twofold screwup issue, and make the parcel logically fragile to estimation mix-ups of RDNN. Hence, NMF achieves a more deplorable introduction than the projected RDNN combinatorial models, particularly in matchless disturbance conditions.

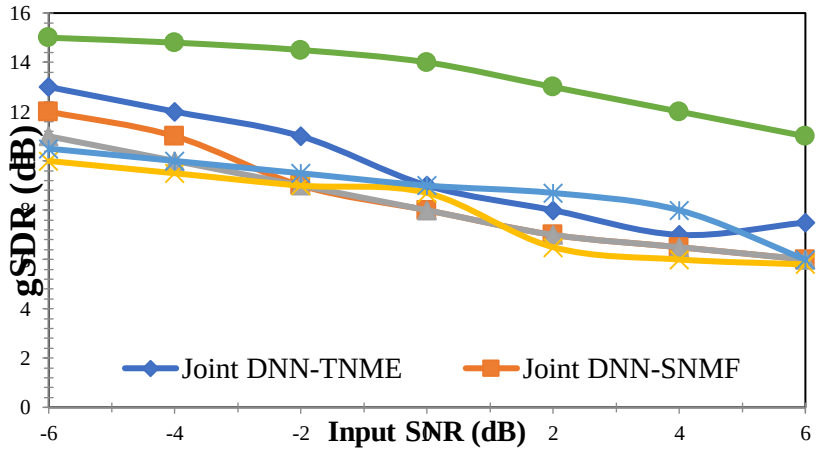


Figure. 4.9: Average gain in SDR: partition execution of a variety of partition prototypes at various input SNR environments

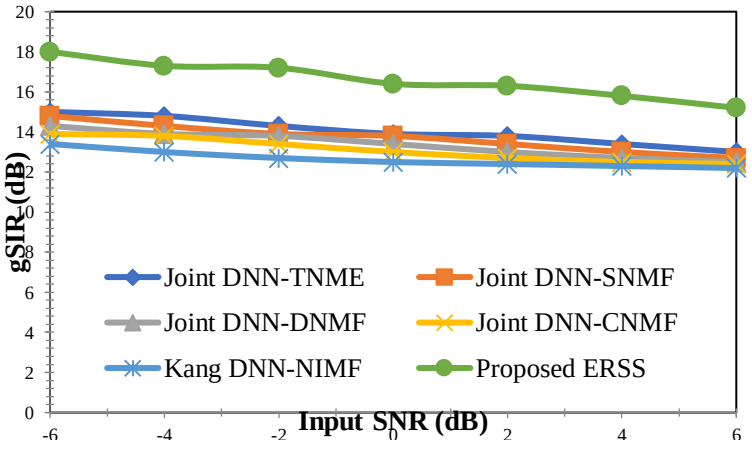


Figure. 4.10: Average gain in SIR: separation performances of a variety of partition prototypes at various input SNR environments

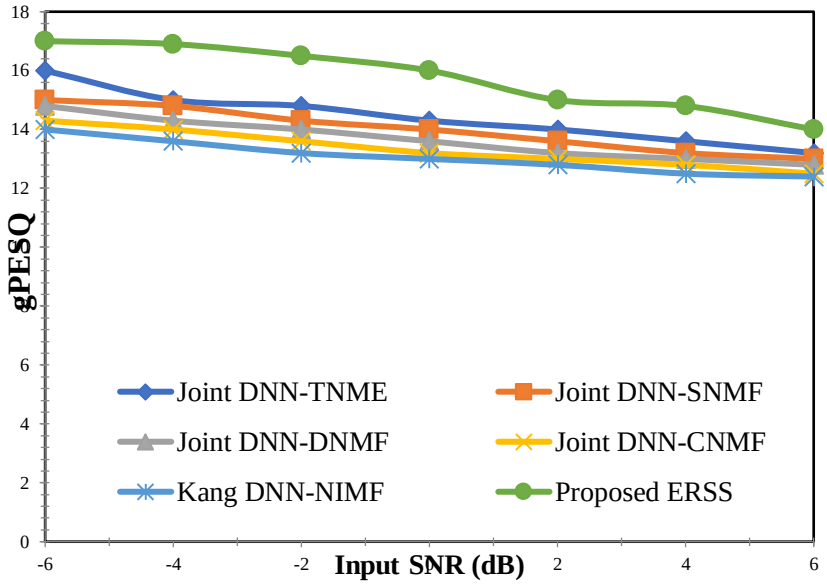


Figure. 4.11: Average gain in PESQ: separation performances of a variety of partition prototypes at various input SNR environments

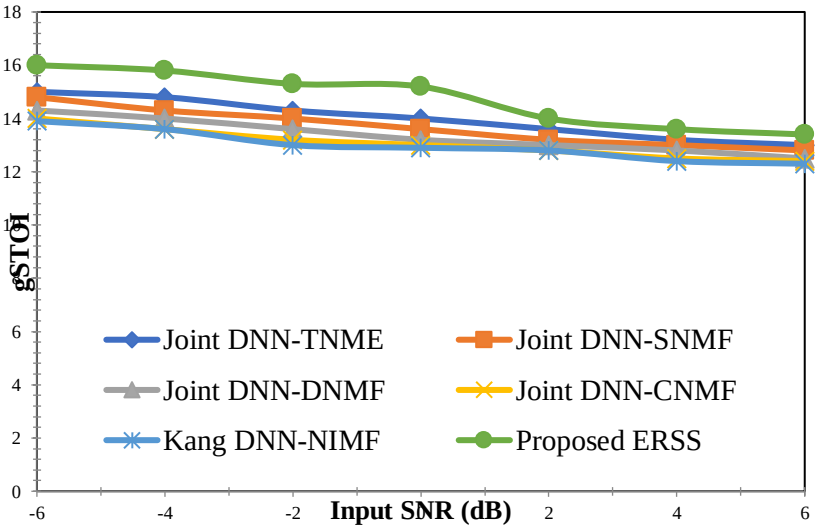


Figure. 4.12: Average gain in STOI: separation performances of various partition prototypes at various input SNR environments

Figures. 4.9 to 4.12 Different performance dimensions show specific and graphical representations of previous technologies. Multiple Frames of Contextual Separation Target You can see that in most evaluations, more than one frame of the separation target is exceeded. This may be why RDNN is best suited to study temporary structures and structural objectives within the separation goal. Compared to individual RTN models, RTN's IFRO's integrated model, DNMF, SNMF, DNMF, and CNMF's integrated model can perform better than speech and speech, so you can learn more about speech and speech.

It is primarily reserved for the joint efforts of DNN and NMF. On the contrary, the spectral-time structure of NMF speech and sound can be applied to the spectrum learned from very pure speech and sound. DNN, on the contrary, has powerful designing technology for non-map learning from input to target. The projected integration plan focuses on the strength of DNN and NMF in sound separation. Kong-TNN-NMF uses both the functions of DNN and NMF for voice partition, but the DNN evaluation of NMF references and functions is done individually or on a tube-by-tube basis. This can lead to double error problems and sensitivity to segregation DNN evaluation errors. Therefore, Kong-DNNNMF has lower performance than the proposed integrated model, especially at unmatched sound levels.

4.6 Summary

In this proposed and completed part of the research work, the data consists of nosy speech signals. The integral fox ride optimization (IFRO) integrated with retrieval-based deep neural network technique has been applied and results have been obtained. It has been observed that the results (performance evaluation metrics) are improved compared with existing works as indicated in the graphs and tables. Experiments show that our proposed method achieves the highest gains in SDR, SIR, SAR STIO, and PESQ outcomes of 10.9, 15.3, 10.8, 0.08, and 0.58, respectively. The Joint-DNN-SNMF obtains 9.6, 13.4, 10.4, 0.07, and 0.50, comparable to the Joint-DNN-SNMF. The proposed result is compared to a different method and some previous work. In comparison to previous research, our proposed methodology yields better results.

Chapter 5

Grasshopper Optimization-based Matrix Factorization (GOMF) and Enthalpy based DNN (EDNN) for multichannel source Separation

5.1 Introduction

For precise signal localization and separation in radar, sonar, and seismology applications, array technologies have been widely used. Since narrowband signals are the major focus of the applications, wideband signals require a generalization of the technology. The directivity pattern of delay-and-sum beamformers is not constant across all wideband signals. Numerous array designs have been suggested as a solution to this issue. Using general design theory as a foundation, Ward presented a constant beam width array. Each microphone signal is treated using a finite impulse response filter in this manner. Filter-and-sum technique is the name given to it. These are reliant on the location of the target, the layout of the array, and they are susceptible to noise signals. A narrowband adaptive beamforming approach was developed for seismic data processing by Capon to address this restriction. Frost developed a wideband adaptive beamformer that was used to filter each microphone signal in an adaptive manner. Griffith and Jim introduced the generalized sidelobe canceller, an enhancement technique. These strategies were well-liked by adaptive beamforming systems.

5.2. Proposed Method

Grasshopper Optimization-based Matrix Factorization (GOMF) and Enthalpy-based DNN are combined to create a novel hybrid approach that is proposed in this chapter for multichannel speech signal separation (EDNN). Before applying the short-term Fourier Transform (SDFT) to the data stream, it is first applied to the multi-channel input signal. An STFT is used to construct complex spectrograms with multiple channel composite waveforms. The fundamental vectors of clean speech are then evaluated using a ranking-based GOMF approach. Then, to distinguish between useful features like directional, spectral, and spatial features, the spatial bearing of the target speaker is used. The spectrogram is then rebuilt using an enthalpy-based deep neural network. Using the inverse STFT (iSTFT) activity, the retrieved yield signal is then transformed back into the produced discourse spectrogram. The proposed strategy's general layout is depicted in Figure 5.1.

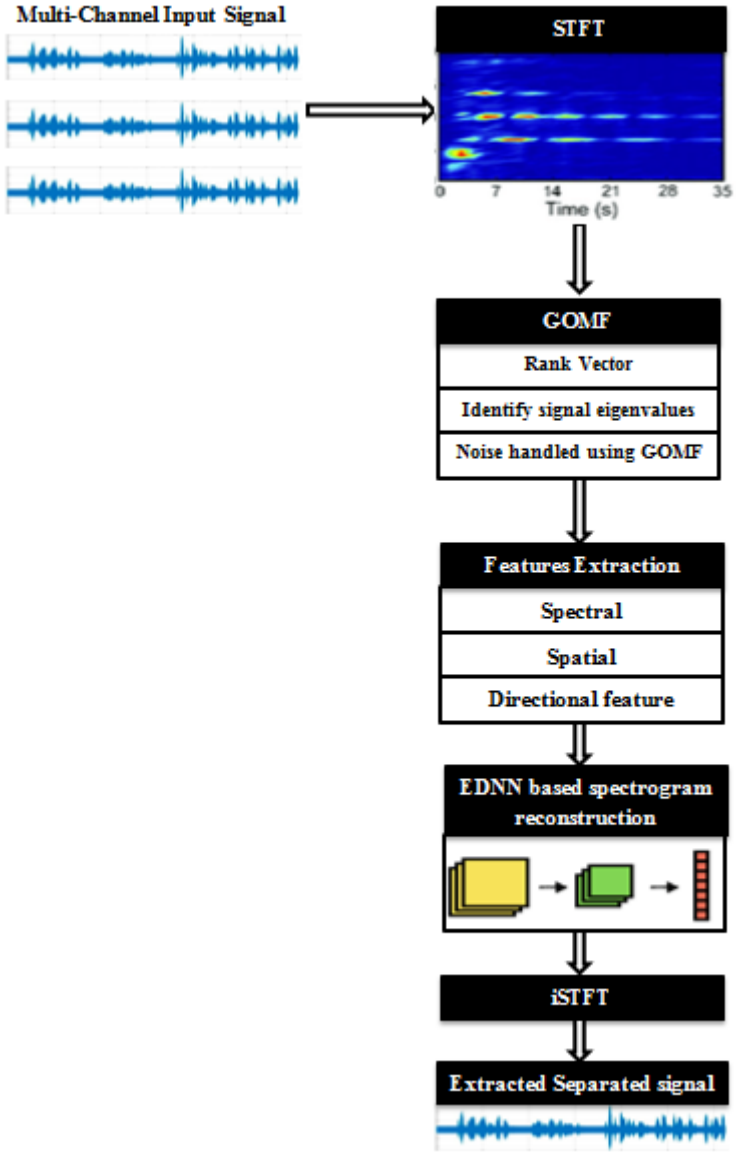


Figure 5.1: Block diagram of the proposed methodology

Figure 5.1 displays the general layout of the model being presented. The multichannel input signal in this instance is put through STFT. In the sections below, the themes are thoroughly explained**;**

5.2.1. Short Term Fourier Transform (STFT)

The multichannel input signal is the first step, followed by a fast Fourier transformation. Complex spectrograms are produced by planning the multi-channel blend waveforms with an STFT [31]. The STFT is a typical Fourier change expansion where the signs are time-varying or non-fixed.

$$Z(y,\ f) = \int z(y_1).h^*(y_1 - y).e^{-(i2\Pi fy)}dy_1 \quad (1)$$

In this instance, the window task is H(y), and the sign is Z(y), both of which are focused at time y. Since the window work has only cut the sign near to time y, the Fourier transformation

is used as a gauge locally at this time. The traditional method of computing the STFT makes use of a fixed positive even window, h(y), of a specific form that is fixated on zero and contains power.

We can create a spectrogram that resembles a conventional Fourier change and range.

$$U_z(y, f) = |Z(y, f)|^2 \quad (5.2)$$

This technique is widely used when examining time-varying and non-fixed indicators. The sign is divided into several smaller components by the spectrogram, and from each component, a range is calculated. This data displays the locations and times of certain frequencies. The multi-channel blend waveforms to complex spectrograms are planned using an STFT. The basis vectors of noise and clear speech are then computed using the rank-based GOMF method. The GOMF idea is thoroughly explained in the following section;

5.3. Based on Grasshopper Optimization, matrix factorization (GOMF)

GOA is relying on the skills and knowledge of grasshoppers. Here, an unique hybrid technique for multichannel speech signal separation is integrated with Grasshopper Optimization-based Matrix Factorization (GOMF). The whole description of the rank estimate process is provided below;

5.3.1. Matrix factorization based on the GOMF:

Step 1: Scov, or spatial covariance matrix

If the scov function is a vector, it will provide a separate power value for each incoming signal, which is presumed to be uncorrelated. If scov is an M-by-M lattice, it speaks to the entire covariance matrix between all incoming signals as shown in the following mathematical expressions:

$$SCM(i, j) = \frac{U(i, j) * N}{U^{'}(i, j) * N} \quad (3)$$

the letters U, I, and V stand for the input spectrum, inverse spectrum, and variance, respectively. Here, U designates the spectrogram input as well as the spectrogram's inverse, while SCM I j) designates the spatial covariance matrix.

Step 2: Initialization

Initialization is a vital stage for the entire optimization process. The multichannel input signals utilised as input in this step are first chosen at random. The length of the grasshopper is N if the overall magnitude of the multichannel input signal is N. The answer's indicator is

grasshopper. The grasshoppers are meant to be arranged in the manner depicted in condition (4), and image 5,2 shows the resolution in action.

$$Grasshopper_i = \begin{bmatrix} G_{11} & G_{12} \ldots\ldots G_{1D} \\ G_{21} & G_{22} \ldots\ldots G_{2D} \\ G_{n1} & G_{n2} \ldots\ldots G_{nD} \end{bmatrix} \quad (4)$$

Initial solution format of OGOA

	S_1	S_2	S_3	S_4		$S_{1000000}$
G_1	1	0	1	0	.....	1
G_2	0	0	1	1	.....	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮
G_n	1	0	1	0	.....	1

Figure 5.2: Solution representation for feature data selection with OGOA

Figure 5.2 illustrates the solution procedure for a sample multichannel input signal. N=1,000,000,000 was used in this work (i.e., number of signals presented in dataset is 1000000). The grasshoppers are shown at random either 0 or 1. This indicates that user data wasn't chosen for the classification procedure at the moment since a grasshopper's station is zero [103]. Otherwise, the data are chosen for the categorization process if the answer is 1. The purchased configuration is offered for the next phase, such as a fitness evaluation.

Step 3: Fitness Calculation

The best rank vector arrangement is then chosen once the fitness function has been looked at. GOMF-controlled noise and signal Eigen values are both recognised by this technique. The fitness function is usually used by optimization algorithms to discover the best configuration. The fitness certification is an essential part of GOMF.

$$Fitness_i = \frac{1}{Noise\,Mean * NoiseVariance} \quad (5)$$

The fitness estimation of each person is evaluated and recorded for future use at the time the first solution and opposing arrangement are developed. Condition is used to illustrate the fitness function (5). We used a multi-target analysis that includes both noise mean and noise variance in this. The use of (6 and 7) allows for the computation of noise mean and noise variance.

$$Noise\,Mean = \frac{sum\,of\,the terms}{number\,of\,terms} \quad (6)$$

$$Noise\,Variance = \frac{\sum (x_i - \bar{x})^2}{n-1} \quad (7)$$

Where "n," the number of observations, "Significance of the Particular Observation," "Mean Value of All Observations," and "n" are pertinent concepts. The solution upgrade process is already ready to proceed with the received results.

Step 4: GOMF based Updating solution

Use the Grasshopper optimization matrix factorization algorithm to adjust the arrangement as needed after assessing your fitness. Using condition, we can modify the solution (8). The grasshopper's circumstance or position can be mathematically described as follows:

$$x_i = S_i + G_i + A_i \ , \quad i = 1,2,...,N \ (8)$$

Anywhere, the ith grasshopper interacts with society in a way called S i, which can be expressed mathematically as follows:

$$S_i = \sum_{j=1, i \neq j}^{N} S(d_{ij}) \hat{d}_{ij}, \quad \mathrm{d_{ij}} = | x_i - x_j | \ (9)$$

The distance between the ith and jth grasshoppers is represented by d ij, whereas s stands for the strength of the social forces function, which can be mathematically expressed as follows.

$$S(y) = fe^{\frac{-y}{l}} - e^{-y} \ (10)$$

Where G I and A I are the gravitational force and wind direction, respectively, for the ith grasshopper, and where the following mathematical equation can be used to express this relationship:

$$G_i = -g \hat{e}_g \ , \quad A_i = u \hat{e}_w \ (11)$$

As opposed to e g and e w, which stand for the unity vector to the centre of the earth and the direction of the wind, respectively, g and u stand for the gravitational constant and constant drift. Nevertheless, equation 11 could not be used directly to determine the solution to the optimization problem, therefore we recast equation 12 as follows:

$$x_i = c\left(\sum_{j=1, i \neq j}^{N} c^{\frac{u-1}{2}} S(| x_j - x_i |) \frac{x_j - x_i}{d_{ij}} \right) + \hat{T} d \ (12)$$

Where, $u \longrightarrow$ higher bound of the search space, $l \longrightarrow$ Inferior bound of the search space,

$Td \longrightarrow$ Best solution value

Step 5: Termination criteria

The optimization procedure comes to an end once the best option has been found. After the rank vector is estimated, features are collected for handling noise and locating signal Eigen values. The parts that follow provide a detailed explanation of the idea that includes NMF.

5.3.2. Normalization mean factorization (NMF)

Using the NMF process, a non-negative grid is declined into two nonnegative lattices, and as

$$UV \approx XY \quad (13)$$

Anywhere one looks, one finds networks with K lines, N sections, and nonnegative components. However, the following noise matrix can be used to describe the NMF model.

$$UV = XY + E \quad (14)$$

calculations aimed at solving the NMF problem and assessing the matrices X and Y derived from the UV objective matrix. They include trading assessment conditions for each lattice.

$$\begin{aligned} X &\leftarrow \arg\min_{X} C(UV \parallel XY) \\ Y &\leftarrow ar\min_{Y} C(UV \parallel XY) \end{aligned} \quad (15)$$

The component in the kth row and rth column of matrix X, which is every member of matrix W, is a distance measure between the matrices A and B given the constraints. There are numerous "distance" measurements that can be used to evaluate C (V || XY), with the Euclidean distance and the Gulbach-Leipler difference being just two examples. The Euclidean distance is used to define C (V || XY) mathematically. $C(UV \parallel XY) = \frac{1}{2} \parallel UV - (XY) \parallel_F^2$ (16)

Where F is the frequent rule and || || The multiplication update rule can be used to reevaluate X and Y as shown below.

$$\begin{aligned} X &\leftarrow X \otimes [(UVY^T)\phi(XYY^T)], \\ Y &\leftarrow Y \otimes [(X^TUV)\phi(X^TXY)], \end{aligned} \quad (17)$$

Anywhere $\otimes and\ \phi$ designate element-wise multiplication as well as division, respectively.

5.3.3. How to calculate the NMF rank:

In this, the fundamental vectors of clear speech and noise are evaluated using a ranking-based GOMF technique. If we rewrite "model" in the following way:

$$\mathrm{UV} = \mathrm{UV}_0 + E \quad (18)$$

Since all matrices X, Y, and = XY have nonnegative rankings, their ranks in the direction of R are all identical. Because of this, the evaluation of a nonnegative position in a noisy environment is inversely correlated with the evaluation of the number of premise vectors. Regrettably, an NP-problematic topic is the surveying of the nonnegative position. Due of the relationship between the rank and the nonnegative position, we prefer to evaluate the rank rather than the nonnegative position (the framework rank is the lower bound of the nonnegative position).

It still requires more than R non-zero Eigen values if we suppose that the rank of is R, even if R = (1/N) only implies R non-zero Eigen values because of X. As a result, the problem with rank evaluation is due to the inability to distinguish between R signal Eigen regards in a noisy environment, which can be solved by noisy head section inspection. The GOMF is a well-known method for analysing model solicitation in the noisy PCA problem; it chooses the model solicitation as the value that limits a threat work. Following rank evaluation and disturbance elimination, the multichannel signal is given to the feature extraction stage, which may be clearly illustrated as follows:

5.4. Extracting Features:

The multichannel signal is shown to highlight the extraction step after rank evaluation and noise removal. The extraction of highlights is a critical step in signal characterisation. Distinguishing a meaningful characteristic from a multichannel signal can be difficult. There are several approaches for extracting elements. In this work, we take multichannel data and extract highlights with extraterrestrial, spatial, and directional bases.

5.5 . Spectrogram reconstruction using an enthalpy-based deep neural network (EDNN):

In contrast to earlier research, which solely relied on DNN, our study uses the Enthalpy algorithm, which we just created and coupled with DNN. This is one of the cutting-edge techniques we used in our research. The convolutional layer of the DNN receives the data initially, which is then subjected to enthalpy, a max pooling layer, a fully connected layer to the Softmax regressor, and a repetition of the procedure.

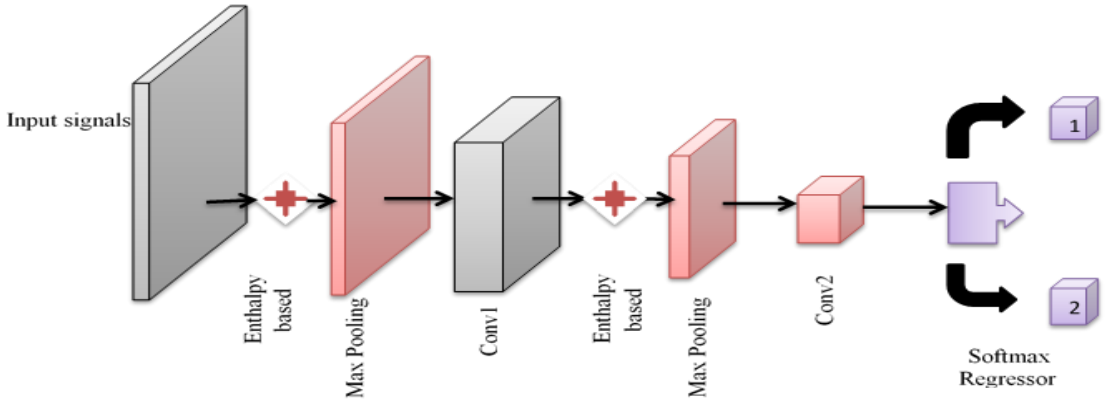


Figure 5.3: Structure of the envisaged EDNN

Figure 3 above depicts the suggested EDNN's design, which is utilised to rebuild the spectrogram. Figure 3 shows the transfer of input signals to the enthalpy layer. Enthalpy starts its process in response to the input signal. In order to reduce the representation of boundaries

and calculations in the system, the max-pooling layer gets the output of the signals after it has completed processing. Before continuing to con1, where it receives the signals with a clear format, Max pooling processes the signals using a matrix or kernel. The signals are then sent back to enthalpy, where they are handled by max-pooling before being transferred to con2. The softmax regressor will receive the full signal from Con2 and forecast the outcome.

5.5.1. Convolution layer

The signals are retrieved in their original clear format by the network's primary layer using a matrix or kernel. Recognizing the pixels aids in maintaining the connection between the signal features. The spectrogram fields should be checked for the upcoming convolutional layer operations. This layer satisfies the criteria set forth by equation (24). In any case, the output of the convolution is referred to as the element map.

$$y_k = \sum_{n=0}^{N-1} x_n h_{k-n} \tag{24}$$

There is usually a filter, input features, and a certain amount of necessities. The output is the yield vector. The subscripts denote the components of the vector.

5.5.2. Layer of normalization depending on enthalpy

Enthalpy-based standardization, the second layer of the network, essentially transmits the indicators to a comparable area or along a specified path. Normalization, which involves altering the signal with the aim of achieving a mean of 0 and a standard deviation of 1, is a common example of a preset go. The process of normalization involves changing the range of pixel force values. The enthalpy-based normalization calculation is carried out to widen the scope of spectrum reconstruction. Using the conditions listed below, a quantitative illustration of enthalpy-based normalization is provided.

$$H = DF + (SEF * SAF) \tag{25}$$

$$EBN = H \times \frac{X}{X_{max} - X_{min}} \tag{26}$$

Where DF stands for direction feature, SEF stands for spectral feature, SAF stands for spatial feature, and combined with are the minimum and maximum values in spectrum reconstruction X. Where EBN is the enthalpy esteem based normalised or spectrum reconstruction.

5.5.3. Max-Pooling layer

The pooling layer reduces the system's representation of computation and boundary. Prior to contributing to the next layer, the max-pooling layer, often referred as as the down sampling layer, is used to reduce the dimensionality of the signal and the yield neurons.

5.5.4. Fully connected layer

Since each neuron receives input from the previous layer, it is advantageous to produce as many neurons as possible from those layers.

Softmax: This term is used to describe the various digits of the labels logistic regression's assumption that there are many classes (0, 1).

$$p_i = \frac{e^{x_i}}{\sum_1^k e^{x_i}} \quad (27)$$

Wherever the network input is, EDNN is employed to categorise the input signals according to the entropy value and to spot abnormal or wounded behaviour. Entropy-based deep neural architecture was implemented in a particular sequence. It also has steps for planning and changing boundary learning.

5.6 Pre-training stage

With the help of the DBN model, the association can produce observable authorizations that reflect its convictions based on the conditions of its hidden units. In this situation, we solved the aforementioned issue using the RBM.

A Restricted Boltzmann Machine (RBM) is a kind of prohibitive Markov self-assertive field with two layers: one layer of stochastic covered (often Bernoulli) units and one layer of stochastic clear (commonly Gaussian or Bernoulli) units. The DNN structure shown in Figure 5.4 demonstrates how it employs a significant number of data neurons to address the selected ideal characteristics and distinctive covered layers before gathering the signals in the yield layer.

Step 1: The observable units, which suggest the selected features to the training vector, are fundamentally introduced.

$$E(x,y) = -\sum_{i=1}^{I}\sum_{j=1}^{J} Q_{ij} f_{si} y_j - \sum_{i=1}^{I} \alpha_i f_{si} - \sum_{j=1}^{J} \beta_j y_j \quad (28)$$

The predisposition word, which indicates the symmetrical collaboration between the detectable component and the concealed component everywhere, describes the number of visible and hidden components. The subordinate log probability of a weight arrangement vector is fundamentally illogical. There are no direct effects between covered units in an RBM, but it is incredibly easy to make a case for

$$\rho(y_j = 1 \mid f_{si}) = \zeta\left(\sum_{i=1}^{I} Q_{ij} f_{si} + \alpha_j\right) \quad (29)$$

Anywhere $\zeta(x)$signifies the strategic sigmoid capacity $\frac{1}{(1+\exp(x))}$, f_{si}, h_j denotes the unbiased sample.

Step 2: In order to make the given visible and undetectable units equal, we update the clear and hidden units. This shows how to conduct the stochastic steepest ascending in the log probability of the arrangement data using a more straightforward learning strategy.

$$W_{ij} = \theta\left(f_{si} y_j\right)_{data} \left(f_{si} y_j\right)_{reconstruction} \quad (30)$$

Once the RBM is ready, a superior RBM can be "stacked" on top of it to produce a multilayer model. communicates at this point about the updated weight as a result of the shifting load in the hidden layer. In the last layer of the correctly arranged layers, a commitment to the novel RBM is secured. Setting up an adjustment stage is the focus of the developed big association burdens.

5.6.1 Fine tuning phase

It is rather common to employ back-spread computation for fine-tuning. To organise system introduction, the DNN is often covered by a yield layer. Similarly, until the advanced weight is mastered or improved, the training dataset is made available. Due to the potential repercussions of missing any indicators along the layout, the DNN classifier is crucial. The classifier in this instance uses the data to complete the procedure. An inverse STFT technique is performed following spectrogram reconstruction. The concept of iSTFT is explained in more detail in the section that follows.

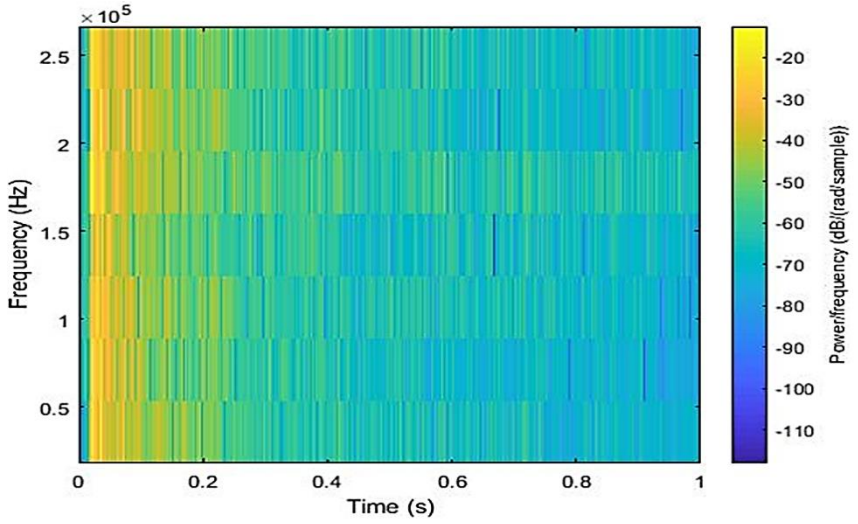
5.7 Inverse STFT (iSTFT) operation

The extracted output signal is then created by converting the generated speech spectrogram using the inverse STFT process. Finally, the extracted separated signal was acquired.

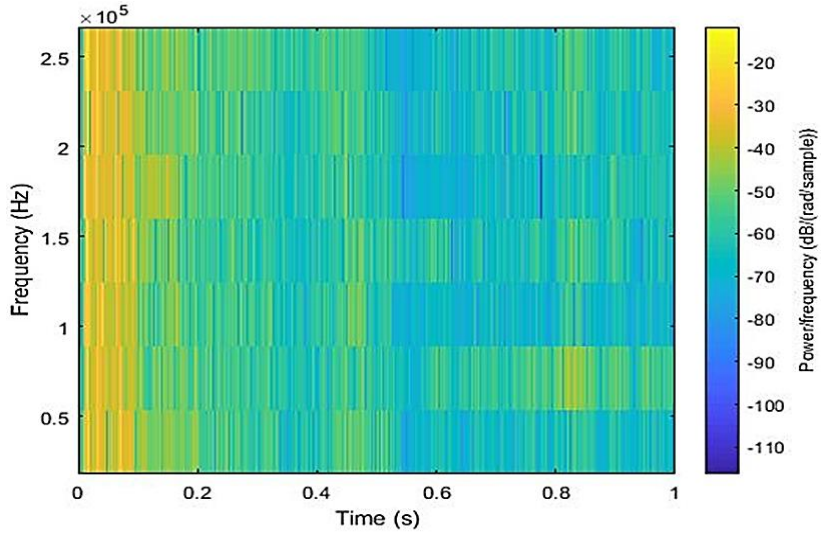
5.8 Results and Conclusions:

The suggested hybrid GOMF and enthalpy-based deep neural network for multichannel speech separation. In this section, the newly proposed methodology is put to use in MATLAB on a system with 6 GB of RAM and an Intel I-7 processor running at 2.6 GHz. Pictures of the iris, which is a distinctive mark, are taken from the dataset and used to evaluate the accuracy and capabilities of the approach.

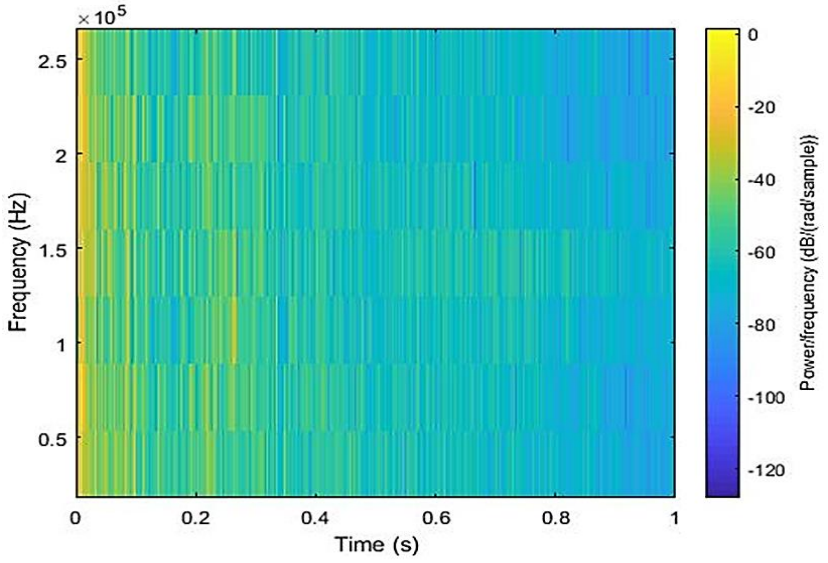
Performance analysis of spectrogram


×10^5
2.5
2
1.5
1
0.5
Frequency (Hz)
0
0.2
0.4
0.6
0.8
1
Time (s)
-20
-30
-40
-50
-60
-70
-80
-90
-100
-110
Power/frequency (dB/(rad/sample))

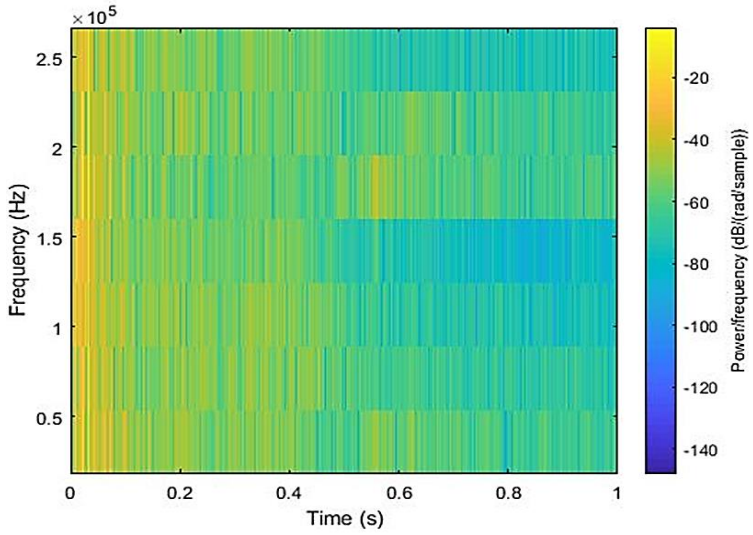
(a)


×10^5
2.5
2
1.5
1
0.5
Frequency (Hz)
0
0.2
0.4
0.6
0.8
1
Time (s)
-20
-30
-40
-50
-60
-70
-80
-90
-100
-110
Power/frequency (dB/(rad/sample))

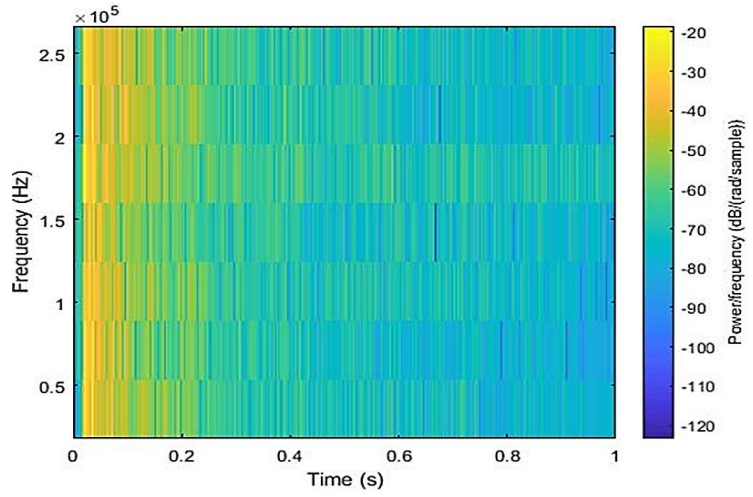
(b)


×10^5
2.5
2
1.5
1
0.5
Frequency (Hz)
0
0.2
0.4
0.6
0.8
1
Time (s)
0
-20
-40
-60
-80
-100
-120
Power/frequency (dB/(rad/sample))

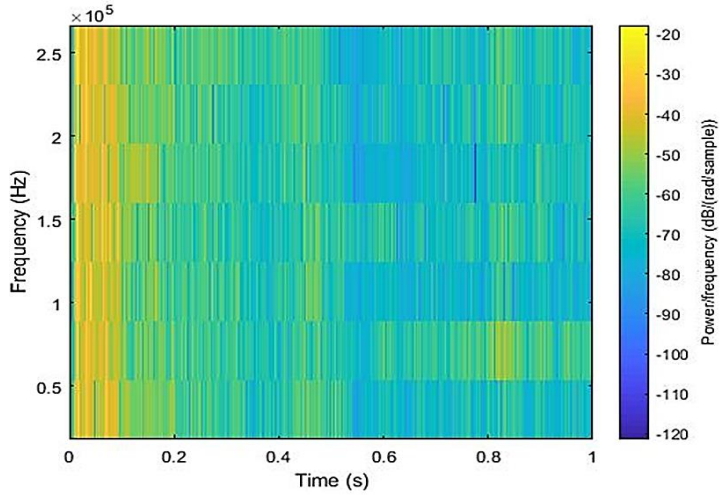
(c)


×10^5
2.5
2
1.5
1
0.5
Frequency (Hz)
0
0.2
0.4
0.6
0.8
1
Time (s)
-20
-40
-60
-80
-100
-120
-140
Power/frequency (dB/(rad/sample))

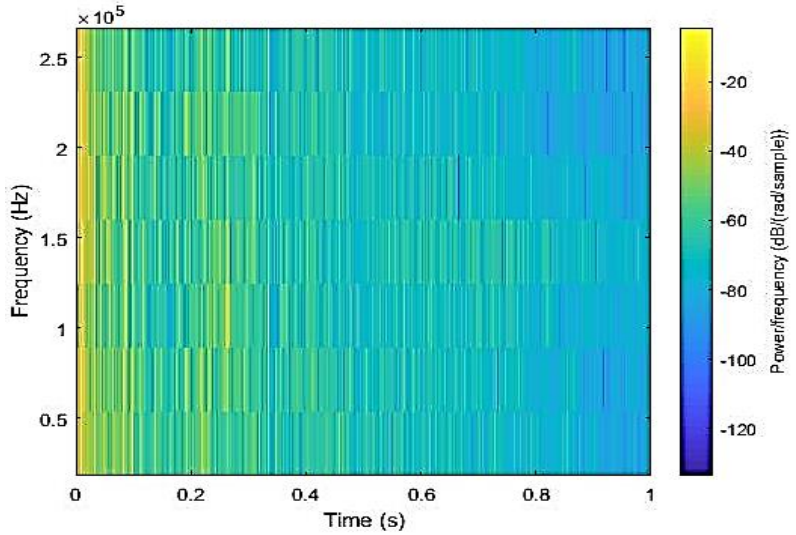
(d)


×10^5
2.5
2
1.5
1
0.5
Frequency (Hz)
0
0.2
0.4
0.6
0.8
1
Time (s)
-20
-30
-40
-50
-60
-70
-80
-90
-100
-110
-120
Power/frequency (dB/(rad/sample))

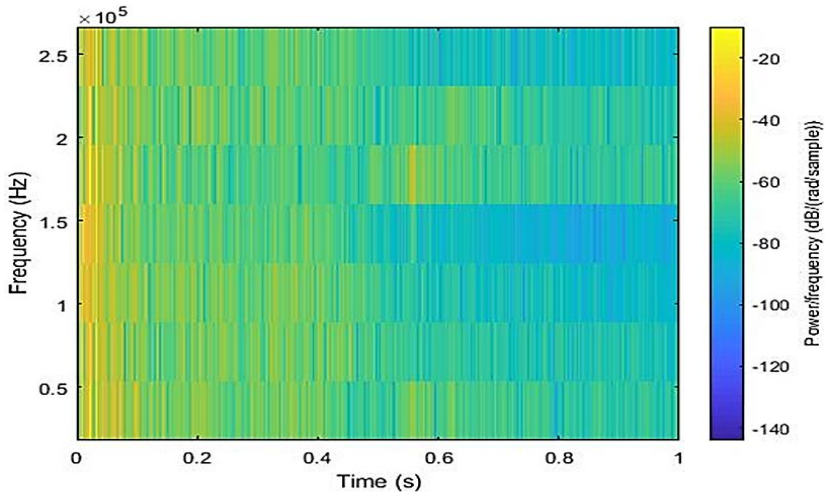
(e)


×10^5
2.5
2
1.5
1
0.5
Frequency (Hz)
0
0.2
0.4
0.6
0.8
1
Time (s)
-20
-30
-40
-50
-60
-70
-80
-90
-100
-110
-120
Power/frequency (dB/(rad/sample))

(f)



(g)



(h)

Figure 5.4: (a), (b), (c), (d) Performance evaluation of the input signal from the spectrogram and (e), (f), (g) (h) Performance evaluation of the signal from spectrogram reconstruction

As a result, Figure 5.4 illustrates the performance analysis of four spectrogram input signals (a, b, c, and d). Four spectrogram reconstruction signals' respective performance analyses are shown in Figure 5.4 (e, f, g, and h). Analysing the aforementioned figures, our suggested changes produce better results. As you can see from the accompanying diagram, our suggested strategy leads to better outcomes than pre-existing theories.

Comparative Results

The system is connected to the most recent DNN-JAT, RNN, and NMF-DNN algorithms. The results are investigated using the SASSEC07 and SiSEC 2010 sets. The (long stretch) SNRs of the signals in the noisy dataset range from - 6 dB to 9 dB. The relative investigation of the current methods is shown in the accompanying tables 1 and 3. Examining voice signals and

signal management strategies is what the speech signal does. As part of the speech processing process, speech signals are gathered, managed, stored, transported, and created. Voice synthesis is the term used to describe the process of creating the information from speech recognition. The four signals SGL1, SGL 2, SGL 3, and SGL 4 are used in this. SGL 1 denotes signal number 1, SGL 2 signal number 2, SGL 3 signal number 3, and SGL 4 signal number 4.

Table 5.1: Comparative evaluation of SASSEC07's data set

Methods	**-6dB**	**-3dB**	**0dB**	**3dB**	**6dB**	**9dB**
PROPOSED	23.1523	23.0523	19.0523	16.0523	11.0523	10.052296
DNN-JAT	17.51032	14.50032	13.50032	10.500317	7.500317	2.500317
RNN	12.5434	11.45434	8.454344	5.454344	2.454344	2.545656
NMF-DNN	11.49991	10.299912	7.299912	5.299912	1.299912	1.700088

In this situation, our recommended approach yields the most extreme result of -6dB of 24.0523, ranging from -6dB, -3dB, 0dB, 3dB, 6dB, and 9dB. According to the analysis, our suggested technique performs better than the present outcomes.

Table 5.2: SASSEC07 Data Set: SDR, SIR, SAR, and PESQ Analysis

	Methods	**SDR**	**SIR**	**SAR**	**PESQ**
SGL1	PROPOSED	65.17269	81.31335	65.28133	4.107441
	DNN-JAT	64.59602	80.84348	64.66359	2.829521
	RNN	63.78988	80.1175	63.89267	1.996385
	NMF-DNN	23.79269	0.746314	21.11269	2.059962
SGL2	PROPOSED	66.5279	84.42081	66.60005	4.03195
	DNN-JAT	64.59602	82.84348	64.66359	2.829521
	RNN	65.05929	82.44476	65.14301	2.14013
	NMF-DNN	22.83533	1.677526	20.56635	1.656663
SGL3	PROPOSED	66.59894	82.61139	66.71209	3.941679
	DNN-JAT	65.72619	81.53611	65.8008	2.817636
	RNN	65.00474	80.92769	65.0953	2.413042
	NMF-DNN	24.36949	2.238257	22.32536	2.117426
SGL4	PROPOSED	68.30567	84.39359	68.42121	4.352382
	DNN-JAT	65.64417	81.72218	65.75984	3.461406
	RNN	64.09938	84.21401	64.14289	2.48984
	NMF-DNN	22.75859	0.754332	20.08968	1.770753

The SASSEC07 data set's SDR, SIR, SAR, and PESQ signal analyses are described in more depth in Table 2 above. In this case, the effectiveness of three various existing methodologies—DMF-DNN, RNN, and DNN—against four different signals is investigated. When analyzing the features in the aforementioned table, our suggestion produces better outcomes. From Table 2 above, it is clear that our suggested strategy outperforms conventional beliefs in terms of effectiveness. Near and to test the suggested Multichannel Speech Separation using crossover GOMF and Enthalpy based Deep Neural Network; this can be sure to confirm the efficacy of the earlier techniques. Figures 5 to 9 show the representation of the spectrum input signals, reconstruction signals, SAR, SDR, SIR, SNR, and PESQ measurements for each dataset.

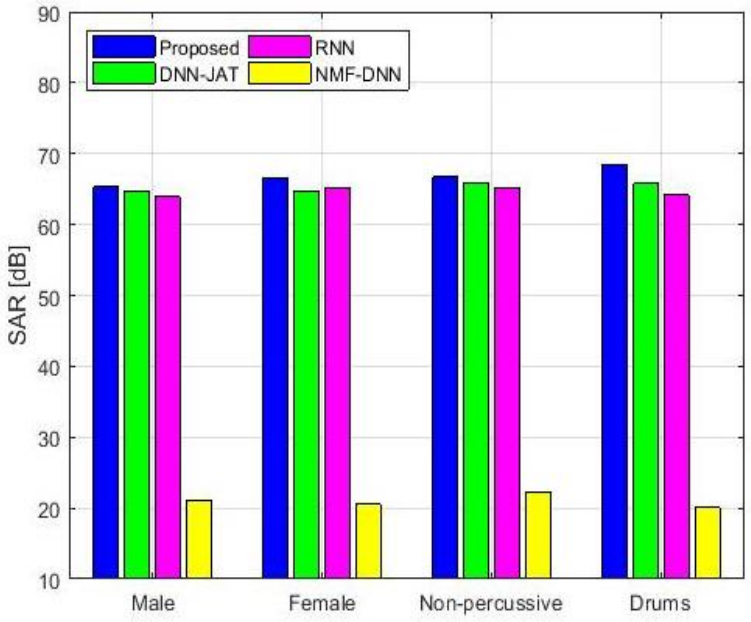


Figure 5.5: SAR performance evaluation

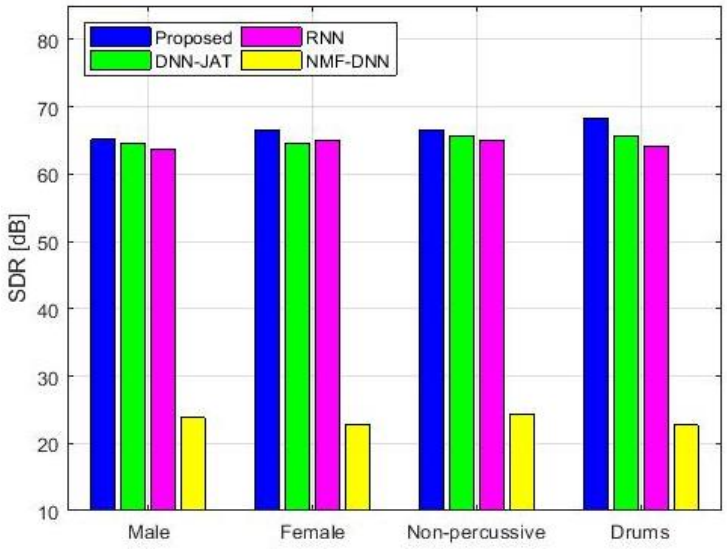


Figure 5.6: SDR performance evaluation

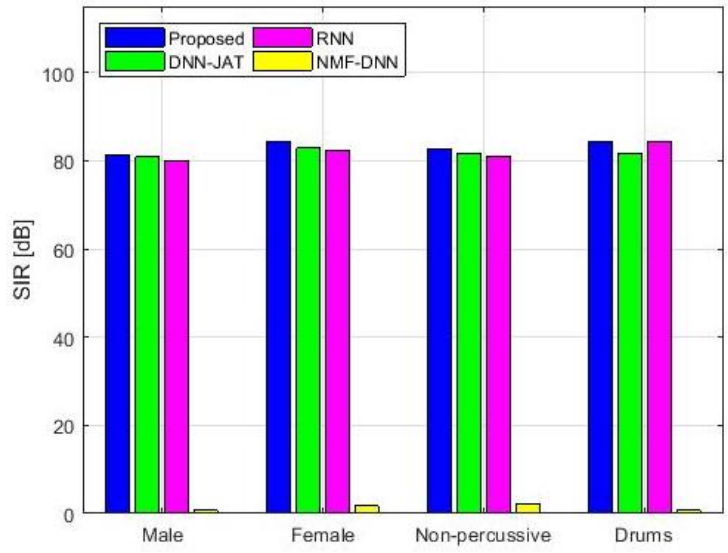


Figure 5.7: SIR's performance evaluation

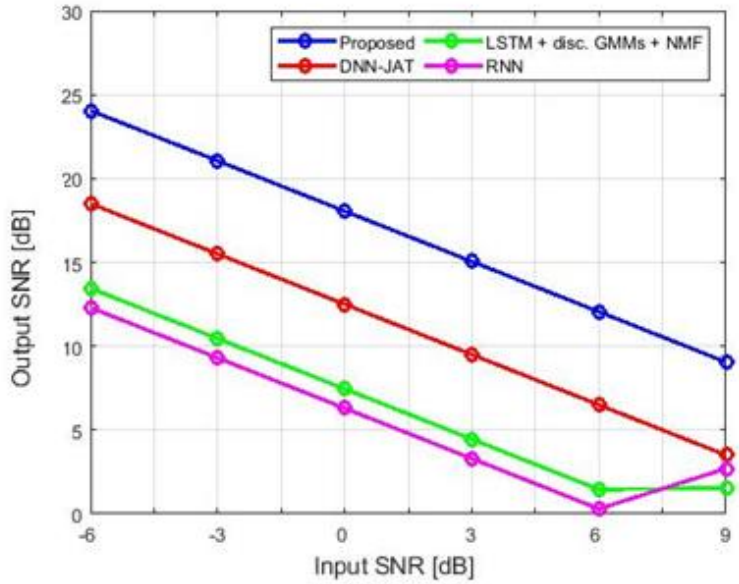


Figure 5.8: SNR performance evaluation

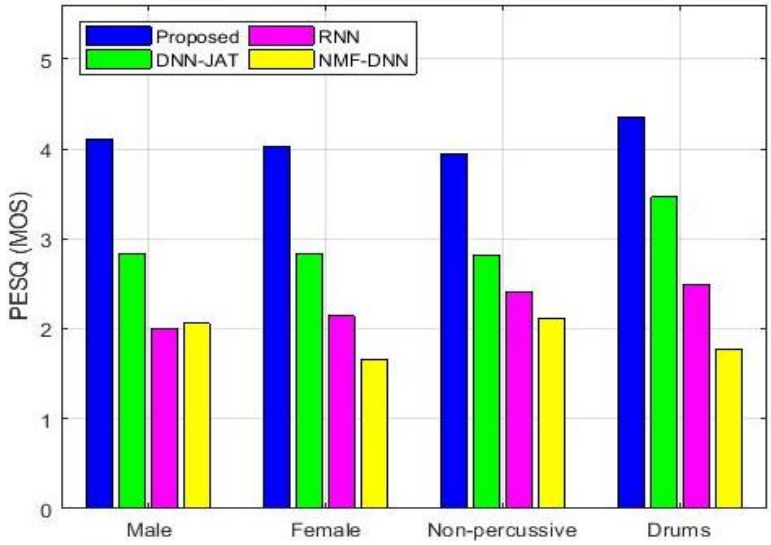


Figure 5.9: Analysis of PESQ score performance

The performance analysis of SAR, SDR, SIR, SNR, and PESQ is shown in Figures 5.5 to 5.9, respectively. PESQ is expressed as mean opinion scores (MOS), with a range of 0 to 5. The higher the MOS, the better. The comparison of the proposed method with the DNN-JAT, RNN, and NMF-DNN existing methods is shown in the above graph. Figure 10-14, when examined, yields the best results in terms of SAR, SDR, SIR, SNR, and PESQ. Our suggested approach yields better results when compared to other current solutions.

Table 5.3: Comparative analysis of data set SiSEC 2010

Methods	**-6 dB**	**-3 dB**	**0 dB**	**3 dB**	**6 dB**	**9 dB**
PROPOSED	22.12368	19.12368	16.12368	13.12368	10.12368	7.123679
DNN-JAT	20.28293	17.28293	14.28293	11.28293	8.28293	5.28293
RNN	17.34285	14.34285	11.34285	8.342846	5.342846	2.342846
NMF-DNN	11.31751	8.31751	5.31751	2.31751	0.68249	3.68249

The presented system attains the extreme outcome of -6dB of 22.12368 ranging from -6dB, -3dB, 0dB, 3dB, 6dB as well as 9dB. From the examination, visibly recognize that the presented technique is enhanced than the current techniques.

Table 5.4: SiSEC 2010_Signal Analysis of SDR, SIR and SAR

	Method	**SDR**	**SIR**	**SAR**	**PESQ**
SGL1	PROPOSED	66.23178	84.69063	66.2942	4.341312
	DNN-JAT	65.77422	82.55409	65.86636	3.048116
	RNN	65.35931	81.53766	65.46658	2.612899
	NMF-DNN	24.6071	0.67683	21.36052	2.089909
SGL2	PROPOSED	66.08751	83.03069	66.18009	3.828596
	DNN-JAT	64.10656	81.23423	64.19774	2.689591
	RNN	65.41014	82.38261	65.50194	2.767913
	NMF-DNN	22.66986	2.860692	20.83926	1.461341
SGL3	PROPOSED	65.83716	83.54766	65.91321	4.0997
	DNN-JAT	65.40494	83.14233	65.47892	3.016585
	RNN	65.39326	78.77577	65.59737	2.566589
	NMF-DNN	22.69306	2.609322	20.78142	1.923716
SGL4	PROPOSED	68.27951	79.65546	68.61231	4.259884
	DNN-JAT	66.39269	78.57407	66.61433	2.974225

RNN	67.92532	77.67461	68.23352	2.621152
NMF-DNN	24.27491	0.697049	20.87947	1.590075

The signal analysis for SDR, SIR, SAR, and PESQ in the SiSEC 2010 data set is described in Table 5.4 above. Four signals are compared in this study using three different existing methodologies: DMF-DNN, RNN, and DNN-JAT. Our suggested modifications result in improved outcomes when compared to the numbers in the aforementioned table. Table 5.4 shows that the proposed technology outperforms the currently employed techniques in terms of results.

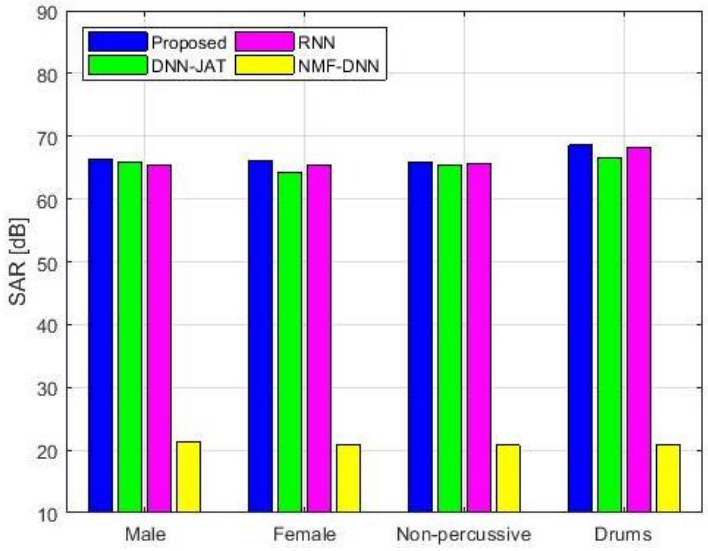


Figure 5. 10: SAR performance evaluation

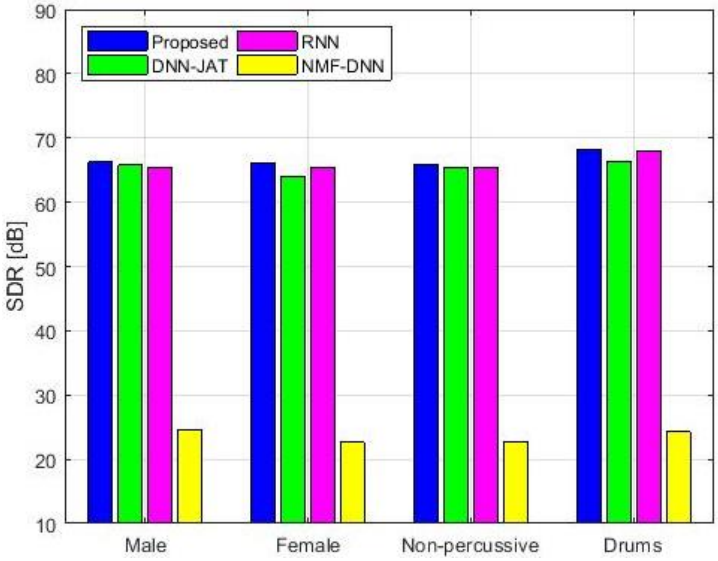


Figure 5.11: SDR performance evaluation

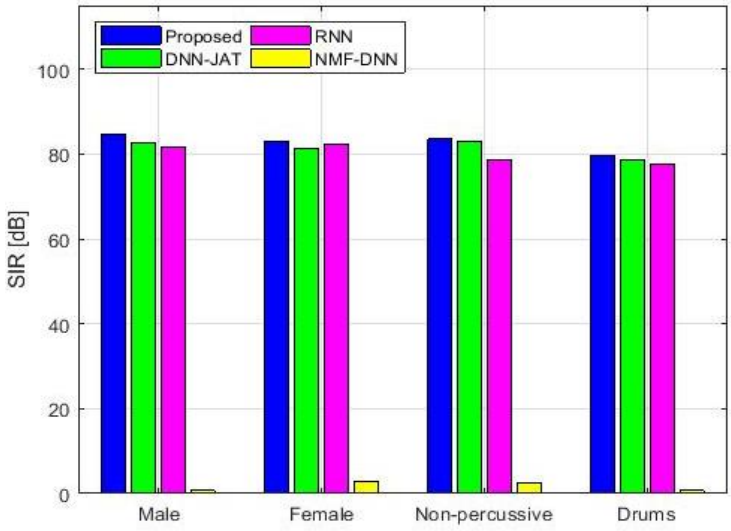


Figure 5.12: Performance review for SIR

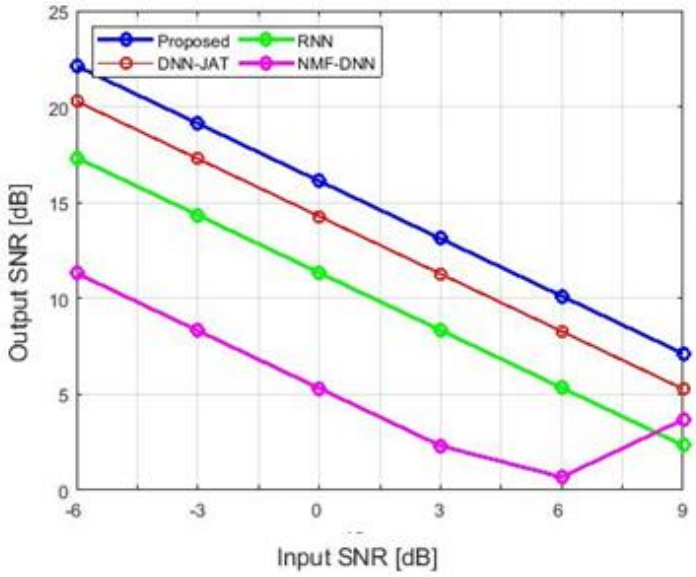


Figure 5.13: SNR performance evaluation

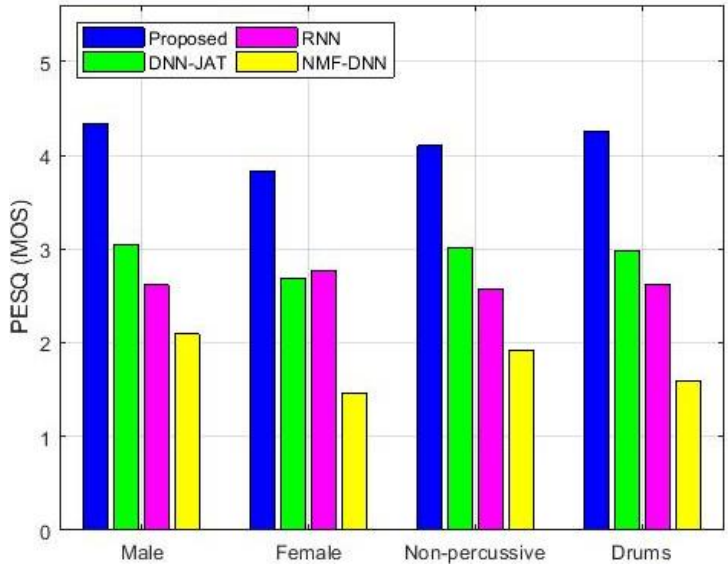


Figure 5.14: Analysis of PESQ Score performance

The graph up top compares the SAR, SDR, SIR, SNR, and PESQ of a proposed approach against those of existing methods using the dataset SiSEC 2010. Figures 10 to 14's analysis shows that the proposed produces the highest SAR, SDR, SIR, and SNR. Compared to other DNN-JAT, RNN, and DNN-NMF techniques currently in use, the one we propose performs better.

5. 9 Summary

Multichannel speech separation is one of the most difficult challenges at the moment. Grasshopper Optimization-based Matrix Factorization (GOMF) and Enthalpy-based DNN are combined and used for data sets (SASSEC07, SiSEC-2010) to obtain results for multichannel source separation. The investigations findings demonstrate better performance when compared with the existing results. It has been observed that the results (performance evaluation metrics) are improved compared with existing works as indicated in the graphs and tables. Experimental results show that our proposed approach accomplishes the most extreme SNR outcome of -6dB of 24.0523. Comparable to the DNN-JAT, which achieves 18.50032. The RNN and NMF-DNN had the worst SNR 13.45434 and 12.29991.

Chapter 6

Krill herd-based matrix factorization (KHMF) and Score-based Convolutional Neural Network (SCNN) for Multichannel Source Separation.

6.1 Introduction

In a previous study, the hybrid Grasshopper Optimization-based Matrix Factorization (GOMF) algorithm shows great potential in the Multichannel speech separation. However, GOMF has a parameter initialization problem and leading to poor separation performance. Instead, a joint creation of the GOMF model parameter approximation and source localization delinquent. So that we proposed Speech Separation with Enthalpy-based DOA and Score-based CNN. The current generation of automatic speech recognition systems can decode clear speech quite well in relatively quiet surroundings, but their performance suffers greatly in loud environments or when a voice signal is present that interferes with the speech signal. Humans, on the other hand, are adept at recognizing combinations of speech signals that are produced by two simultaneous speakers.

There are many techniques that have been developed to enhance voice recognition in the presence of background noise or competing speech. Among them, the methods (1) multichannel signal separation, also known as blind signal separation, and (2) computational auditory scene analysis may show promise (CASA).

6.2. PROPOSED SYSTEM

This chapter suggests a technique for decoding multi-channel speech signals that combines enthalpy-based DOA, KHMF, and score-based CNN. Determine the signal's STFT first. The branch begins the subsequent phase by determining the enthalpy of the signal under analysis. The change in space energy caused by DOA in each time interval is referred to as enthalpy. The spatial energy histogram will be transformed by the GMM that determines the enthalpy function at each time frame. The SCM model is parameterized by the enthalpy DOA in the third step using the signal tracker's output as a basis. To calculate the tracked address, use multi-channel KHMF. In the fourth step, useful features are extracted that correspond to the spatial direction of the target speaker, such as directional features and spatial features. A SCNN ratio based on the score will then be used to mask the spectrogram. In the method of forming, the photo blocks are visible.

In contrast, no specific prior knowledge is necessary for the multichannel signal separation technique. It only makes use of statistical data from the multivariate data collected from a collection of microphones, where variations in propagation delay might be significant. An automatic speech recognizer can use the multichannel signal separation as a front-end to separate the simultaneous speech signals into individual signals, thereby cancelling the cross-talk for a particular speech signal. As a result, it is anticipated that it will enhance the target speech source's recognition performance, as current automatic speech recognition systems are very sensitive to cross-talk and perform significantly worse in this environment than they do in environments with other types of background noise.

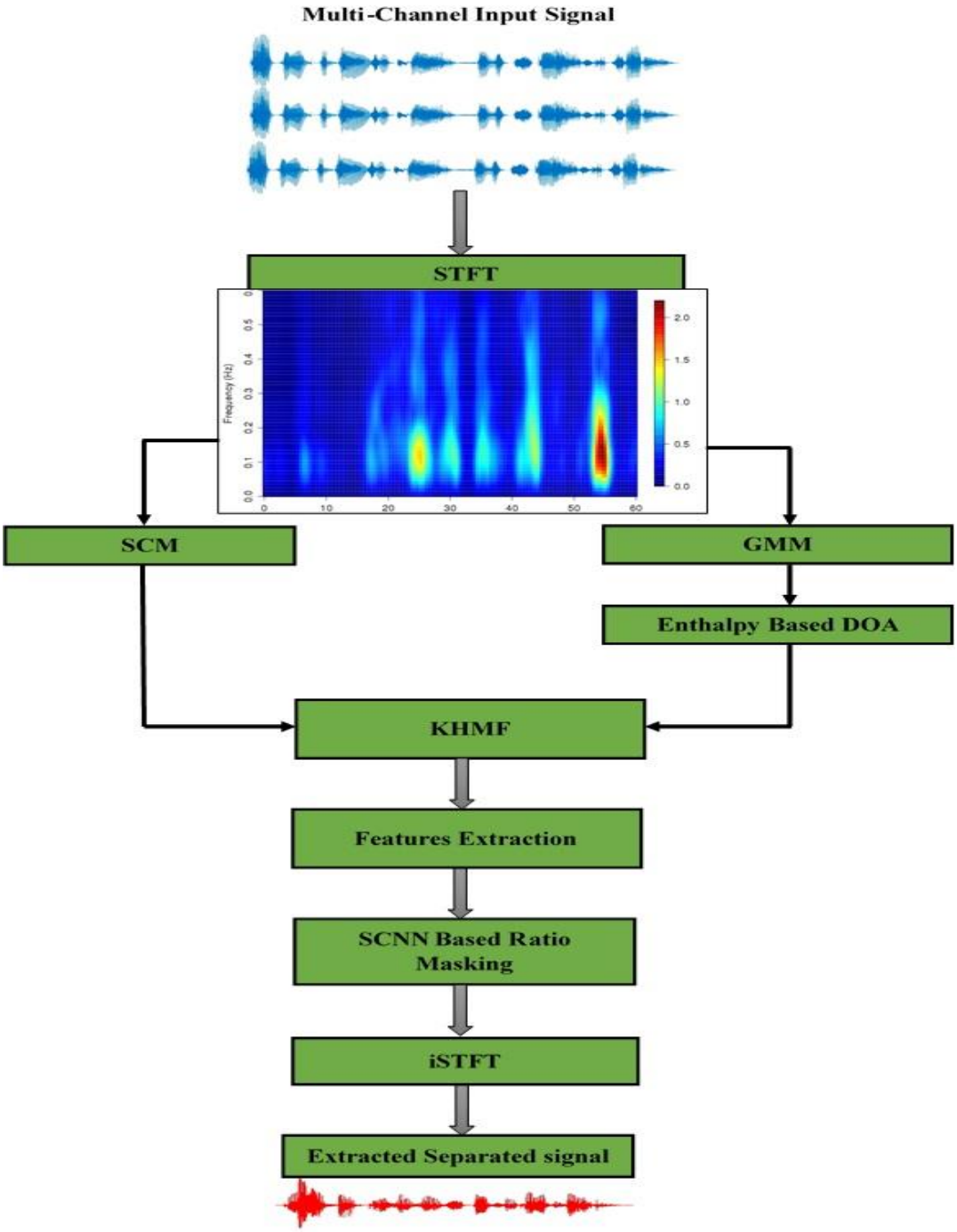


Figure 6.1: shows a block diagram of the suggested approach.

The offered model's schematic is shown in Figure 1. Multi-channel input signals are employed in this case with STFT. These ideas are thoroughly explained in the following sections.

6.2.1. STFT

Multichannel input signals are received, and then the short-term Fourier transform is applied. Through the use of STFT, complex spectrograms are created from multi-channel mixed waveforms. The STFT, a distinct extension of the Fourier transform, is utilised when the signs are variable or not fixed.

$$Z(y, f) = \int z(y_1).h^*(y_1 - y).e^{-(i2\Pi fy)}dy_1 \quad (1)$$

Anywhere, Z(y) denotes the symbol, and h(y) denotes where the temporary work is located within the y window. The Fourier shift serves as a local indicator at time y, the only time the window truncates the sign. A fixed positive even window h(y), which needs to be zeroed out and given a base, must be used in some way to calculate STFT. We can design the spectrogram as using the Fourier transform and normal range.

$$U_z(y, f) = |Z(y, f)|^2 \quad (2)$$

It is typically used to examine signals that evolve over time. The spectrogram separates the sign into numerous smaller parts and estimates the range of each part, giving us knowledge of the moment when several frequencies converge. In intricate spectrograms, it is used to plan multi-channel mixed signals. The monitoring branch then begins by calculating the analysis signal's enthalpy. Enthalpy describes how space energy changes with DOA in each time interval. The spatial energy histogram is transformed into DOA measurements by the GMM, which calculates the enthalpy function at each time frame. The definition of granularity is provided in the following paragraph.

6.2.2 GMM:

The goal of the GMM [23] is to identify the mixture that most accurately reproduces the multivariate Gaussian probability distribution of an input set. In this scenario, each time interval will estimate the Gaussian model of the mixture of enthalpy that transforms the space energy.

To model the spatial distribution of the mixture, we advise utilising a mixture model as opposed to searching for SRP peaks [24]. For each time frame of the guided response performance, the Gaussian value was assessed independently (SRP). A DOA measurement value (multiple directions of arrival) with mean, variance, and weight is created using the GMM result parameter from the discrete spatial distribution acquired from the SRP. Sound hopping across borders is the source of the noise in SRP. The use of GMM can lessen the effects of noise if the measurement uncertainty in multi-channel speech separation can be represented by the width of each peak provided by the Gaussian variance from e to h.

The probability density function (PDF) of univariate Gaussian distribution [25] [26] through mean μ as well as variance σ^2 can be defined as follows;

$$\begin{aligned} S(\theta;\mu,\sigma^2) &= \sum_{i=\infty}^{\infty} Z(\theta;\mu+i2\pi,\sigma^2) \\ &= \sum_{i=-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} u^{-\frac{(\theta-\mu+2\pi i)^2}{2\sigma^2}}, \end{aligned} \tag{3}$$

Where $N(\theta;\mu,\sigma^2)$ is a PDF of a regular Gaussian distribution, l is the GMM index of 2π multiples, and $\theta \in [-\pi,\pi]$. Here, the GMM through weights a_k designed for every Gaussian distribution k is well-defined as algorithm1;

Algorithm 6.1: EM-Algorithm for estimation of GMM

Input: Signal data S

Output: μ

Initialize the a, μ and σ^2

Compute probability density function using equation(3)

//First get the equation in E step

For *t=1:T*

For *i=1:Z*

For *k=1:K*

$$\eta d_{ki} = \frac{S(\theta_{ki};\mu_{ki}+2\pi,\sigma_{ki}^2)a_{ki}}{\sum_k^K = \sum_i^\infty = S(\theta_{ki};\mu_{ki}+2\pi,\sigma_{ki}^2)a_{ki}}$$

End

End

//Second go through the M step

For *d=1:D*

For *k=1:K*

$$\mu_{dk} = \frac{\sum_d^M = \sum_k^\infty = \eta d_{dk}(\theta_{dk}-2\pi)}{\sum_d^M = \sum_k^\infty = \eta d_{dk}}$$

$$\sigma_{dk}^2 = \frac{\sum_d^M = \sum_k^\infty = \eta d_{dk}(\theta_{dk}-\mu_{dk})}{\sum_d^M = \sum_k^\infty = \eta d_{dk}}$$

$$a_{dk} = \sum_{d}^{M} = \sum_{k}^{\infty} = \eta d_{dk}$$

End

End

Until converge reached the condition.

The above algorithm specifies the estimation as well as maximization of the GMM.

$$S(\theta; a, \mu, \sigma^2) = \sum_{k=1}^{k} a_k \sum_{i=-\infty}^{\infty} S(\theta; \mu k + i2\pi, \sigma_k^2) \tag{4}$$

Anywhere k in the entire quantity of Gaussians in the model as well as EM procedure for approximating parameters $\{a, \mu, \sigma^2\}$ that exploit the log-likelihood

$$\log L = \sum_{d=1}^{M} \log \sum_{k=1}^{k} a_k \sum_{i=-\infty}^{\infty} S(\theta_d; \mu_k + i2\pi, \sigma_k^2) \tag{5}$$

Is assumed in [25]. The parameter θ_d indicates the standpoints of guidelines indices $d = 1,..., M$ used to estimate SRP in (24).

6.2.3 Enthalpy based DOA:

The spatial energy histogram is then translated into DOA measurements in each time interval using an estimated GMM of the enthalpy function. This can be written mathematically as:

$$EBTF = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \tag{6}$$

$$Ep = EBTF + \mu_k \tag{7}$$

Anywhere, $X_{\min}$ along with $X_{\max}$ are the minimum as well as maximum values into c measurements X, where EBTF is the enthalpy-based time frame.

6.2.4 DOA Measurements by GMM:

For both time frames, Algorithm 1 operates discretely. In addition, the resulting means via variances and weights are assessed as permuted DOA measurements. A combination of Gaussians aiming at both time frames is obtained. The algorithm is currently unable to distinguish between measurements in each frame that are brought on by real sources and those that are due to noise. Enthalpy is used to explain the connection between the DOA and the space energy of each period in this context. The spatial energy histogram is transformed into DOA measurements at each time frame by the estimation of the enthalpy function. A spatial covariance matrix model (SCM model), which is parameterized by DOA based on enthalpy, is created once the DOA measurement results have been calculated. According to the signal, the tracker outputs. In order to estimate the spectral model from the source in the direction described in more detail below, use the multi-channel KHMF to represent the spatial behaviour of the source in time;

6.2.5 Spatial Covariance Matrix Model:

The signal tracker output is used to define a spatial covariance matrix model [27] parameterized by Enthalpy-based DOA. For each time frame point in this example, the SCM is calculated. Both input channels' magnitude spectrograms are contained in each diagonal. The disagreement and absolute value of (off-diagonal values) denote, respectively, the segment variance and magnitude correlation among microphones for a time frame point. Combination SCMs can be used to approximate the TF domain mixing in equations (2, 3).

$$M_{xy} \approx \hat{M}_{xy} = \sum_{s=1}^{S} H_{xy,s^q} {}_{xy,s} Ep \tag{8}$$

Where the source's optimistically valued magnitude spectrogram and the frequency domain Room Impulse Response (RIR) SCMs are. Then, using the multi-channel KHMF to estimate the spectral model of the source from the tracking direction as will be detailed below, the obtained SCM represents the spatial behaviour of the source in time;

6.2.6 KHMF:

By modelling the grazing of krill populations based on certain organic and ecological forms, KHA is a new meta-heuristic technique that can rationalise the population to address the reproduction problem. This optimization algorithm's goal is to maximise herd density. Here, matrix factorization is the foundation of the conventional krill swarm optimization approach. The flow representation based on the krill population's matrix decomposition is depicted in Figure 6.2. In the part after this, this procedure will be explained.

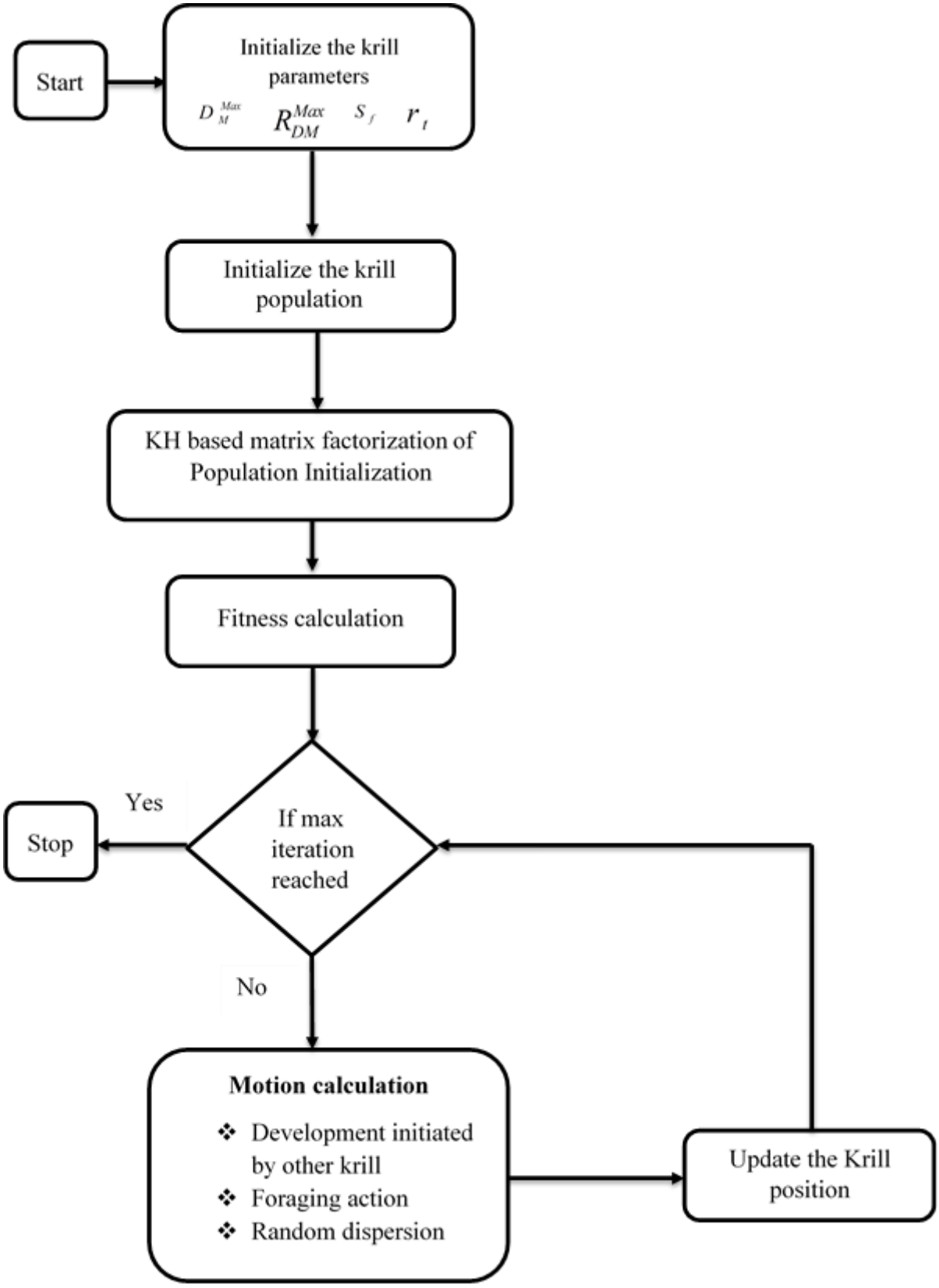


Figure.6.2 The flowchart for krill herd-based matrix factorization algorithm

The next section will introduce the step-by-step process of the matrix factorization algorithm based on krill swarms.

6.3 Krill Herd based Matrix factorization:

Step 1: Spatial covariance matrix (SCM):

Each input signal receives a separate power value if scov is a vector [28], which is likewise regarded as unimportant. If an M-by-M network, then the full covariance matrix involving all input signals is being discussed. The following is its mathematical expression:

$$SCM(i,j) = \frac{U(i,j) * N}{U^{'}(i,j) * N} \tag{9}$$

Anywhere; $U \rightarrow$ Multichannel input signals, $U^{'} \rightarrow$ Inverse multichannel signals, N $\rightarrow$variance value. Here, *SCM (i, j)* stipulates the spatial covariance matrix, *U* stipulates the multichannel input as well as stipulates the inverse of signals.

Step 2: Initialization

The population size, overall evolution number, and are the key KHA factors. The feature value is represented by the krill herd in our proposed method. We obtain some sets of initial solutions after initialising the values. The following steps receive these solutions.

Step 3: Fitness calculation

Estimate the fitness effectiveness rest on the equation (10) and select the best result.

$$Fitness = \max PSNR \tag{10}$$

Krill herd-based matrix factorization repeats the application of the first three movements while also adhering to the search directives to increase the value of the goal function. Every individual krill's mobility is controlled by three key mechanisms.

(a) Development initiated by other krill individuals,

(b) Foraging action,

(c) Random dispersion.

6.3.1 Development initiated by other krill individuals

Individual krill attempt to maintain increased thickness throughout this process, while the rate of development of other krill affects the speed of each individual. To evaluate the motion-induced effect, three impacts are used: the repulsive impact (x), the neighbourhood impact (y), and the objective impact (z). For each individual m, this signal could be expressed as

$$D_m^{new} = \xi_m D^{\max} + \chi_b D_m^{old} \tag{11}$$

Where,

$$\xi_m = \xi_m^{current} + \xi_m^{target} \tag{12}$$

$$\xi_m^{current} = \sum_{n=1}^{N} F_{mn} P_{mn} \tag{13}$$

$$P_{mn} = \frac{P_m - P_n}{abs(P_n - P_m) + rand} \tag{14}$$

$$F_{mn} = \frac{F_m - F_n}{F^w - F^b} \tag{15}$$

$$\xi_m^{target} = K^{best} F_m^{best} P_m^{bset} \tag{16}$$

$$K^{best} = 2\left(random + \frac{M}{M_{\max}} \right) \tag{17}$$

$$\xi_m = \sum_{n=1}^{N} \left[\frac{F_m - F_n}{F^w - F^b} \times \frac{P_m - P_n}{abs(P_n - P_m) + rand} \right] + 2\left(random + \frac{M}{M_{\max}} \right) F_m^{best} P_m^{best} \tag{18}$$

$D^{\max}$ - Extreme induced signal or motion, χ_b - Inertia weight of the motion-induced inside the range [0, 1], D_m^{old} - Preceding induced motion of the m^{th} krill individuals, $F^w and F^b$ -The most horrible also the finest situation amid altogether the krill individuals of the population, $P_m and P_n$ -Current situation of the m^{th} as well as the n^{th} entities, N - Amount of krill individuals additional than the specific krill, $M and M_{\max}$ - Amount of present iteration in addition to an extreme quantity of iterations, F_m^{best} -The best fitness value of the m^{th} and the n^{th} individuals, P_m^{best} -The best-related position of the m^{th} and the n^{th} individuals.

At this point, a parameter termed as sensing signal distance S_d is utilized for the distance amongst the individual krills as well as the neighbours also it is expressed by,

$$S_d = \frac{1}{5N} \sum_{n=1}^{N-1} \left| F_m - F_n \right| \tag{19}$$

Where N - Entire amount of the krill individual, $F_m - F_n$ -Position of the m^{th} as well as n^{th} krill.

6.3.2 Foraging action

This action is founded upon dual foremost factors. Originally the current food area, as well as the second, is the data about the previous food area. For the m^{th} krill individual, the foraging velocity can be spoken by,

$$F_{Fm}^{new} = S_f \varsigma_m + \chi_x F_{Fm}^{old} \tag{20}$$

Where, χ_x -Inertia weight of the foraging motion, F_{Fm}^{new} *and* F_{Fm}^{old} -Foraging motions of the new and the old m^{th} krill

6.3.4 Random dispersion

To improve the populace variety random diffusion procedure is mostly measured as well as it is spoken by,

$$R_{Dm}^{new} = \beta \times R_D^{\max} \tag{21}$$

Where, $R_D^{\max}$ - Maximum diffusion speed, β - Random directional vector lies amid [-1, 1].

Step 4: Updating the position

In this process, a single krill can potentially shift from its current position to one that is more beneficial because to the random movement of induction movement, feed movement, and propagation. The promoted placement of the mth krill individuals throughout t and also may be associated by, as shown by the three investigated movements above.

$$P_m(t + \Delta t) = P_m(t) + \Delta t \frac{dP_m}{dt} \tag{22}$$

Where, An n-dimensional judgment space in the Lagrangian model is used to express basic KHA technique as shown below,

$$\frac{dP_m}{dt} = D_m^{new} + F_{Fm}^{new} + R_{Dm}^{new} \tag{23}$$

Where D_m^{new} -the motion-induced through additional krill individuals, F_{Fm}^{new} - foraging motion, R_{Dm}^{new} -physical diffusion of the krill individuals

$$\Delta t = r_t \sum_{n=1}^{N} (UL_n - LL_n) \tag{24}$$

Where UL_n *and* LL_n -Upper and lower limits, r_t -Random number uniformly distributed between 0 to 2. Based on the above method, SCM represents the spatial behaviour of the source in time and uses multi-channel KHMF to estimate the spectral model from the source following the direction.

Step 5: Termination criteria

After finding the best solution, the optimization process ends. After the evaluation, the result will be passed to the feature extraction step, which will be explained in detail below.

6.4 Feature Extraction:

At this point, actual features like directivity and spatial features are extracted based on the speaker's spatial direction. Next, a convolutional neural network model estimated from the spectrogram is used to mask it.

6.4.1 Feature extraction based on the directional feature (DF):

In this instance, we separate the target speaker using a neural spatial filter, a direct function in a neural tissue model. The two-layer directional highlight can be effectively planned and then incorporated into the creation of the conventional multi-channel voice segment in accordance with the previous characteristics of the solid support. To get the target speaker ready for separation, emphasise at the information level (for instance, the power spectrum and the space allocated between channels).

Here, two new directions—directional power ratio (DPR) and directional signal-to-noise ratio—are applied in consideration of the operating capacity of the universal fixed strip forming machine (DSNR). Some static channels, such as the super cardioids static pillar before, are intended and predetermined by way of, which imagine recovering sound sources as course for reappearance canister. These channels are directed at, and receiver demonstration and a pre-characterized bearing lattice are assumed. We can use the organising produce intensity of as a functional assessment of the significant force from course given that we anticipate that these immovable channels can stretch entirely about spatial detachment and that the numerous speakers are not firmly located in the space. As the marker is a T-F receptacle overcome through the sign from heading, the DPR can be calculated in this way and is classified as follows:

$$DPR_{\theta p}(t, f) = \frac{\left\| W_p^H(f) Y(t, f) \right\|_2^2}{\sum_{k=1}^{P} \left\| W_k^H(f) Y(t, f) \right\|_2^2} \tag{25}$$

Somewhere; $Y(t, f)$ is the polyhedral spectrum vector in bin $TF(t, f)$. In addition, in most radiation pattern design methods, each fixed spatial filter has multiple rejection regions. For example, the signals near θ_{P} are well preserved by $w_p(f)$, but they are greatly attenuated by $w_k(f)$, $\theta_k \in \Omega_p$. At this point, Ω_p is a set of directions, and its radiation pattern in the θ_{P} direction is zero. It can be precisely defined during the design phase of the beam-former. If the address grid covers the entire space, DSNR can therefore be interpreted as the ratio of signal power θ_{P} to the strongest interference:

$$DPR_{\theta p}(t,f)=\frac{\left\|W_p^H(f)Y(t,f)\right\|_2^2}{\max\limits_{k\in\Omega_p}\left(\left\|W_k^H(f)Y(t,f)\right\|_2^2\right)} \tag{26}$$

In this case, the directivity of DPR and DSNR can provide clues to distinguish the target speech from the interference.

6.4.2 Feature extraction based on the spatial feature (SAF):

We initially use cochlear gramme decomposition to separate the left and right ear signals from among these spatial features [29]. In particular, a 64-channel gamma-ray channel that controls a register with a focus frequency between 50 Hz and 8000 Hz by means of a proportionate rectangular transaction rate scale divides the information mix. Each channel's power is restricted to half-wave support, a track motion of 10 ms, and loop lengths of 20 ms. We omitted the two primary binaural accents of ITD and ILD because the TF nameplate has a sampling rate of 16 kHz and 320 models can be enumerated by binaural information prompts. ITD is built on a common CCF between the left and right hemispheres, as the Lyr application obliquely suggests.

$$CCF(c,m,\tau)=\frac{\sum_k x_{cm,l}(k)x_{cm,r}(k-\tau)}{\sqrt{\sum_k x^2cm,l^{(k)}}\sqrt{\sum_k x^2cm,r^{(k-\tau)}}} \tag{27}$$

XCM,L and XCM,R transfer the symbols of one side of the device and the right ear in channel c and freely group m under the predetermined parameters, and k records a sign instance of a T-F unit. is between -1 ms and 1 ms. The CCF component for a test frequency of 16 kHz is 33. To examine characters coming from various starting points, the CCF aspect is used as a partial vector right away.

Another two-dimensional (2D) ITD will be implemented at this point. The CCF estimate at the anticipated delay e in relation to the target speech head serves as the primary measurement. The result is the highest CCF score, which measures the compatibility of the left and right hearing aids and is used to decide which binaural decorations to use to reduce noise. The usage provides suggestions for these two highlighted objectives. To identify directional sources of scattered noise, the highly regarded CCF is employed. The least common CCF value should be close to 1 for directional sound sources and close to 0 for diffuse sound sources. The evaluation target's discrete speech and annoying noise, which are brought on by another source, are directly resolved by taking CCF into account.

$$ITD(c,m)=\begin{pmatrix}CCF(c,m,\tau\\ \max CCF(c,m,\tau)\end{pmatrix} \tag{28}$$

In specific ILD associates towards the energy ratio in DP, and is determined under every unit pair,

$$ILD(c,m) = 10\log_{10}\frac{\sum k^{x^2} cm, l^{(k)}}{\sum k^{x^2} cm, r^{(k)}} \tag{29}$$

Overall, we may say that 2D ITD and 1D ILD make up each pair of TF blocks' space allocation. To create a spatial component vector in the envelope, we connect each projection at the unit level. The overall measurement value for each time window for a 64-channel cochlea is 192. The extracted features will proceed to the following SCNN step after feature extraction. Next, based on the neural network masking factor, convolution is used to estimate the spectrogram. The following provides a thorough explanation of the convolutional network concept.

6.5 SCNN:

Three layers, including a convolutional layer, a clustering layer, and a fully linked layer, make up the proposed SCNN. The weights and biases of the preceding layer influence the CNN classifier's final judgement. The condition (26) and condition (27) of each layer, in turn, justify these weights and biases.

$$\Delta W_l = -\frac{x\lambda}{r}W_n - \frac{x}{N_t}\frac{\partial C}{\partial W_n} + m\Delta W_n(t) \tag{30}$$

$$\Delta B_n = -\frac{x}{n}\frac{\partial C}{\partial B_n} + m\Delta B_n(t) \tag{31}$$

Where W_n represents the weight, B_n represents the bias, n signifies the layer number, λ signifies the regularization parameter, x represents the learning rate, N_t represents the total number of training samples, m represents the momentum, t represents the updating step, and C represents the cost function. The CNN classifier includes various kinds of layers are as follows,

(a) Convolutional layer: It contains several learned weighting matrices, so-called filters, which slide on the input signal. In each convolutional layer, the performance of the transmission layer is first checked according to various learning weight networks called template filters. Operate linearly to output the layer. This layer uses condition (30) to perform convolution of the input data and the kernel. The result of convolution is also called an attribute map.

$$C_k = \sum_{m=0}^{M-1} y_n \hat{h}_{k-n} \tag{32}$$

Anywhere, y_n is the input features $\hat{h}$ is the filter and M is the number of components in y and the output vector is C_k.

(b) Pooling layer: This layer is called the down sampling layer. The clustering process reduces the size of neurons emerging from the convolutional layer to diminish computational intensity as well as avoid over-fitting. In this sense, the largest grouping activity will select the most excessive stimulation in each component. Reduce the number of output neurons. In addition, the grouping layer shortens the information in the output of the convolutional layer.

(c) Fully connected layer: This level is completely related to each start of the previous level. That is, this layer connects each neuron in the maximum combination layer with all output neurons. The activation function used in this work corresponds to the following:

Softmax: This function calculates the probability distribution of k output categories. Therefore, the output layer uses the softmax function to calculate the input category corresponding to normal or abnormal.

$$p_i = \frac{e^{x_i}}{\sum_1^k e^{x_i}} \tag{33}$$

Anywhere, χ is the multichannel input signals that are, the output classes of SCNN are extracted output signal. After restoring the spectrogram, perform the reverse STFT operation. The next section will introduce the concept of iSTFT in detail. STFT inverse operation.

6.5.1 Inverse STFT (iSTFT) operation:

Finally, the STFT inverse operation is applied and used to modify the resulting speech spectrogram. At the end of the extracted output signal, we get the extracted single signal.

6.6 Results and Conclusions:

The proposed multi-channel KHMF is used for speech separation using enthalpy-based DOA and SCNN. In this section, the method introduced in MATLAB applies to a system with 6 GB RAM and an Intel I-7 processor. The accuracy and performance of the method were evaluated at 2.6 GHz, and Signals were collected from the data set.

6.6.1 General Assumptions:

In the experiment, we used 50 professional music recording datasets from SiSEC 2018 [30]. Here, clear language and diffuse noise are selected from the TIMIT corpus [31]. To test the common p-dimensional situation, we used three real-time voice mixing and three

microphones and 4 signal sources from SiSEC 2011 [32], level 3×5 (3 mixed signals-5 signal sources), and Level 4×8 (4 mixed signals-8 signal sources).) Random male and female voices. In the 3×5 example, we mixed 5 audio sources, and in the 4×8 example, we mixed 8 audio sources. Since there is no reliable information about the angle at which the source is placed in the mix, we use our proposed method to estimate the DOA of the source.

6.6.2 Comparative Results:

In order to analyse the results using SiSEC 2018 and the TIMIT suite, the proposed system makes use of existing Directional Fuzzy C-Means (DFCM), Weighted Mixture of Directional Laplacian Distributions (WMDLD), Flexible Audio Source Separation Toolbox (FASST), and GaussSep algorithm (GS) methods. The comparative analysis of the suggested and existing methodologies is shown in the accompanying tables 6.1 and 6.2. Each data set in this instance has two blends. Male2, Male3, and Male3 are created from these two combinations, as indicated below.

6.7 Comparative Analysis of Dataset SiSEC2018:

Mixture Signal 1:

Table 6.1: Mixture1 Data set SiSEC2018 Analysis of SDR, SIR, and SAR

	Methods	SDR	SIR	SAR
Male2	**PROPOSED**	**45.53**	**35.16**	**10.37**
	DFCM	06.54	15.81	07.43
	WMDLD	06.51	17.62	07.29
	FASST	06.18	10.86	08.44
	GS	10.83	16.56	12.32
Male3	**PROPOSED**	**27.31**	**37.88**	**10.57**
	DFCM	04.14	12.25	05.37
	WMDLD	04.07	13.77	05.12
	FASST	03.25	07.91	06.35
	GS	05.64	11.53	07.29
Male4	**PROPOSED**	**45.30**	**35.45**	**09.85**
	DFCM	09.39	18.55	10.06
	WMDLD	08.69	19.78	09.21
	FASST	07.63	10.98	10.89
	GS	13.18	20.23	14.19

Mixture Signal 2:

Table 6.2: Mixture2 Data set SiSEC2018 Analysis of SDR, SIR, and SAR

	Methods	SDR	SIR	SAR
Male2	**PROPOSED**	**25.92**	**33.88**	**07.97**
	DFCM	06.46	14.94	07.35
	WMDLD	06.57	16.92	07.22
	FASST	04.38	07.32	08.85
	GS	10.83	16.56	12.32
Male3	**PROPOSED**	**28.43**	**34.56**	**06.08**
	DFCM	08.23	16.35	09.16
	WMDLD	07.80	17.7	08.81
	FASST	07.57	11.91	13.37
	GS	10.79	16.82	12.45
Male4	**PROPOSED**	**25.95**	**34.52**	**08.58**
	DFCM	08.03	14.81	05.75
	WMDLD	05.77	16.61	08.81
	FASST	04.66	11.91	13.37
	GS	07.57	16.82	12.45

Tables 6.1 and 6.2 above display the analysis of SDR, SIR, and SAR signals in the SiSEC 2018 data set. In this case, the multi-channel signal is compared to some state-of-the-art methods, such as DFCM, WMDLD, FASST, and GS. For each set of data, the input signal with the best outcome and the signal for spectrum reconstruction are shown below;

Figures 6.3 to 6.12 depict, respectively, the performance analysis of SAR, SDR, and SIR. The above picture illustrates the strategies using the current DFCM, WMDLD, FASST, and GS techniques. Numbers 3 to 12 are anticipated to have the highest SAR, SDR, and SIR results. Our concept outperforms alternative approaches in terms of results.

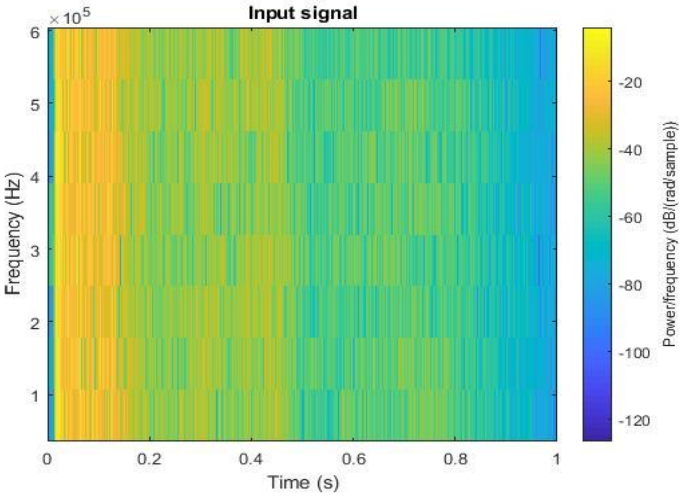


Figure 6.3: Performance analysis of input signal SiSEC Mix1

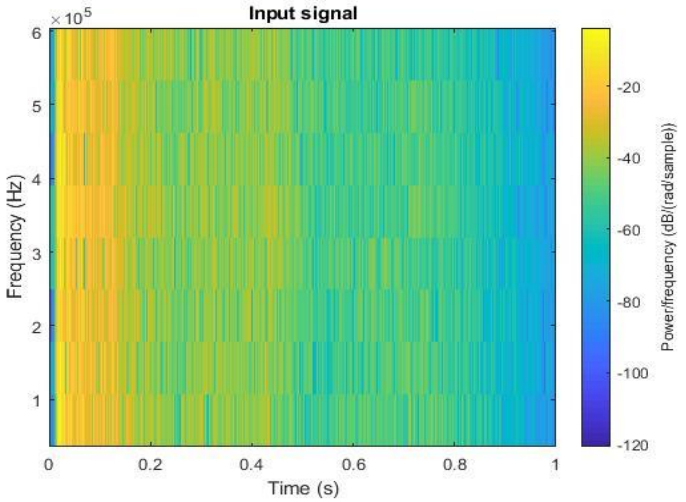


Figure 6.4: Performance analysis of input signal SiSEC Mix2

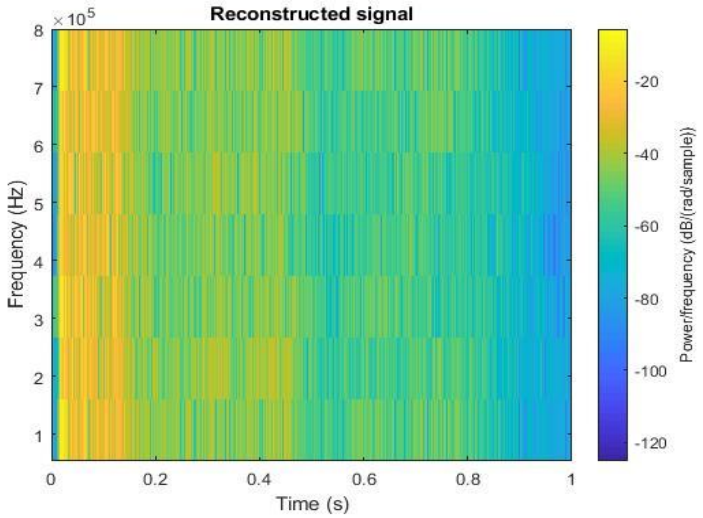


Figure 6.5: Performance analysis of Reconst signal SiSEC Mix1

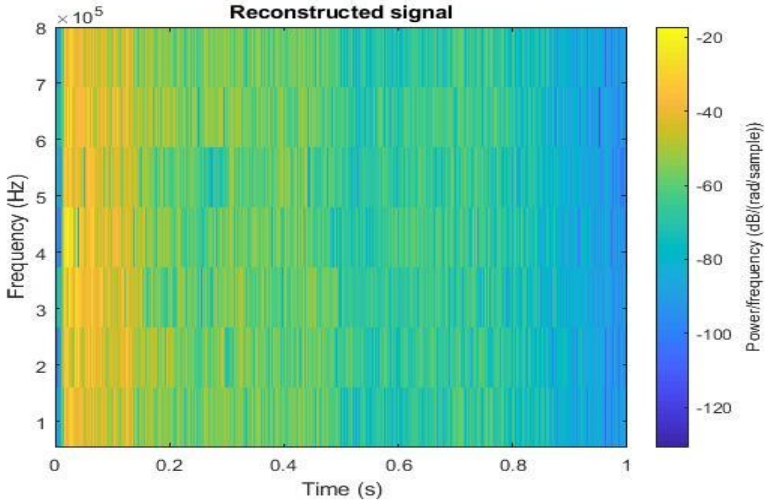


Figure 6.6: Performance analysis of Reconst signal SiSEC Mix2

Figures 6.3 to 6.12 depict, respectively, the performance analysis of SAR, SDR, and SIR. The above picture illustrates the strategies using the current DFCM, WMDLD, FASST, and GS techniques. Numbers 6.3 to 6.12 are anticipated to have the highest SAR, SDR, and SIR results. Our concept outperforms alternative approaches in terms of results.

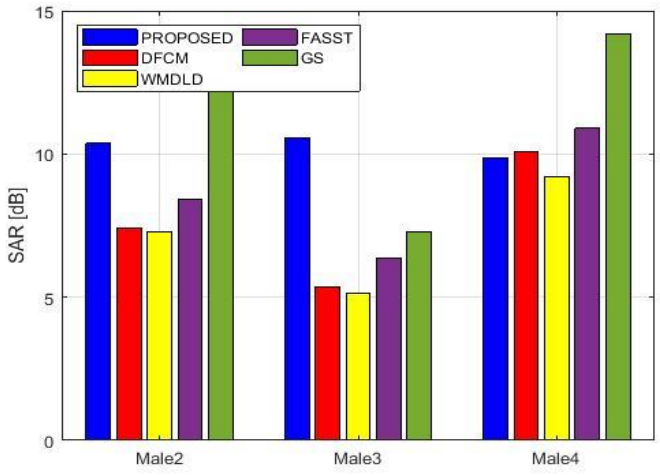


Figure 6.7: A comparative analysis of SAR Mix1

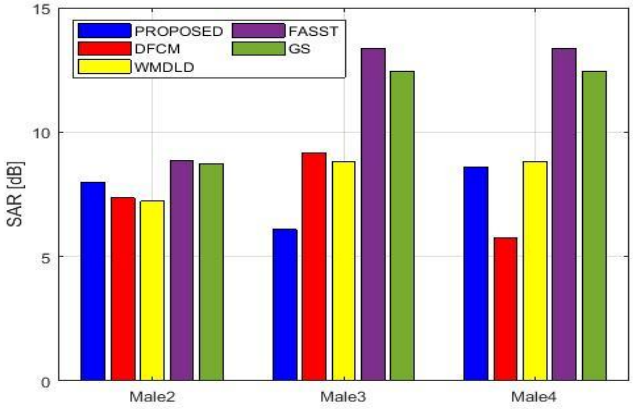


Figure 6.8: Comparative analysis of SAR Mix2

The comparison between the DFCM, WMDLD, FASST, and GS methods and the SAR mixes 1 and 2 is made from the aforementioned figures 6.7 and 6.8. The source-to-artifact ratio (SAR) gauges a network's ability to provide extraordinary superiority results without adding more artefacts. Analysis of the aforementioned figures 7 and 8 reveals that our solution yields more favourable outcomes.

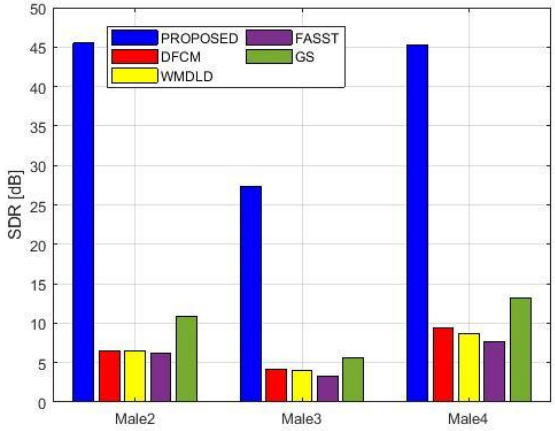


Figure 6.9: comparative analysis of SDR Mix1

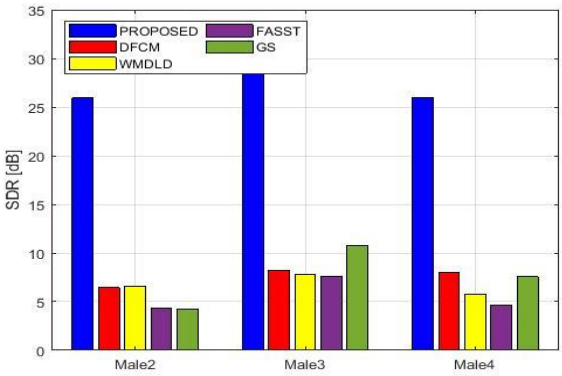


Figure 6.10: Comparative analysis of SDR Mix2

In the image above, the existing DFCM, WMDLD, FASST, and GS methods are contrasted with the 6.9 and 6.10 SDR mixture1 and mixture2. SDR is widely used in this context as a general indicator of a source's audio quality. The most cutting-edge time- and frequency-domain resolution is at odds with SDR (source-to-distortion ratio). In comparison to figures 6.9 and 6.10 above, our proposed yields superior outcomes.

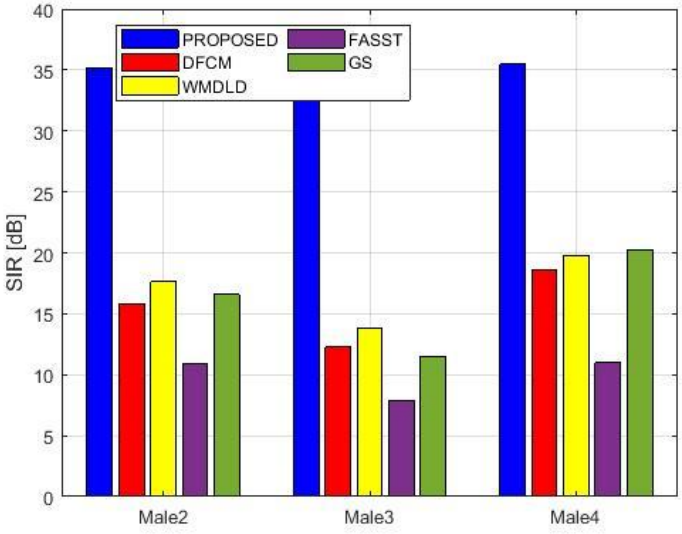


Figure 6.11: a comparative analysis of SIR Mix1

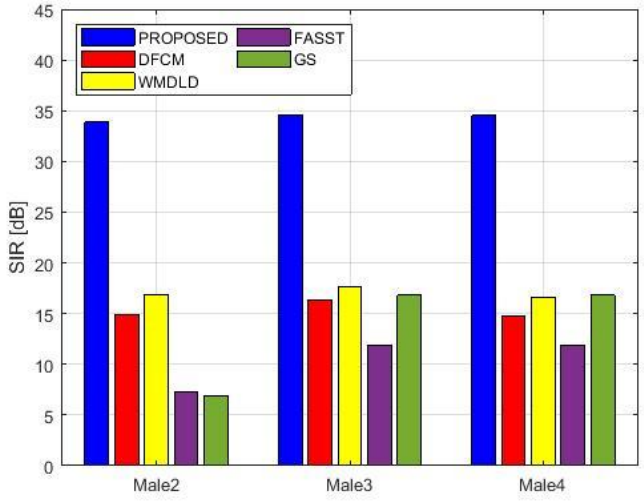


Figure 6.12: a comparative analysis of SIR Mix2

Using the aforementioned figures 6.11 and 6.12, SIR mixture1 and mixture2 are compared to the present DFCM, WMDLD, FASST, and GS methods. The source of interference ratio (SIR), as a result, is a statistic that demonstrates how well the algorithm can

maintain the source of interest while eliminating other sources. Analysis of the aforementioned figures 6.11 and 6.12 shows that our solution yields more favourable outcomes.

6.8 Comparative Analysis of Dataset TIMIT:

The current approaches CTF-MINT, CTF-MPDR, BP, and Unproc are connected to the suggested system. Analyse the outcomes using the TIMIT kit. An analysis of the new method and the old method is shown in the attached figure. according to the following;

Cepstral distance (CD) (dB), frequency-weighted segment SNR (FWSeg.SNR) (dB), and log-likelihood ratio (log-LR) were the three evaluation criteria specified in REVERB Challenge [27]. (LLR). In order to assess the voice source's performance in terms of separation, we also employed the SDR (dB) and SIR (dB) in [34]. The difference between the estimated value of the single reverberation signal and DE and the estimated value of the microphone input signal should be calculated for each measurement. The lower the score, the better when employing CD and LLR. In other words, a higher score is preferable. Results evaluation based on the 1 and 2 metre distances are shown in Table 3 and Table 6.4.

Table 6.3. Evaluation results: distance is 1 m

Method	SDR diff	SIR diff	CID diff	FWSeg.SNR	LLR
Proposed	**8.1**	**10.22**	**-0.54**	**2.32**	**-0.19**
CTF-MINT	8.05	10.18	-0.53	2.29	-0.19
CTF-MPDR	7.71	10.02	-0.5	2.14	-0.17
CTF-BP	7.4	9.58	-0.47	2.09	-0.16
Unproc	5.71	5.77	-0.25	0.8	-0.12

Table 6.4. Evaluation results: distance is 2 m

Method	SDR diff	SIR diff	CID diff	FWSeg.SNR	LLR
Proposed	**7.54**	**8.67**	**-0.41**	**1.62**	**-0.18**
CTF-MINT	7.48	8.62	-0.40	1.61	-0.18
CTF-MPDR	7.22	8.46	-0.38	1.51	-0.16
CTF-BP	6.79	8.02	-0.35	1.41	-0.16
Unproc	5.55	4.25	-0.16	0.6	-0.11

Figures 6.13 to 6.21 illustrate, in turn, how the TIMIT dataset's comparative analysis of SDR-based SNR, SIR-based SNR, PESQ-based SNR, TIMIT reconstruction signal SNR, SIR-based NPM, and PESQ-based NPM is analysed using the current CTF-MINT, CTF-MPDR, BP, and Unproc techniques. Analysis of figures 6.13 to 6.21 reveals that CTF-MINT, CTF-MPDR, BP, and Unproc have the highest gains. In relation to additional common approaches, our suggested achieves better results.

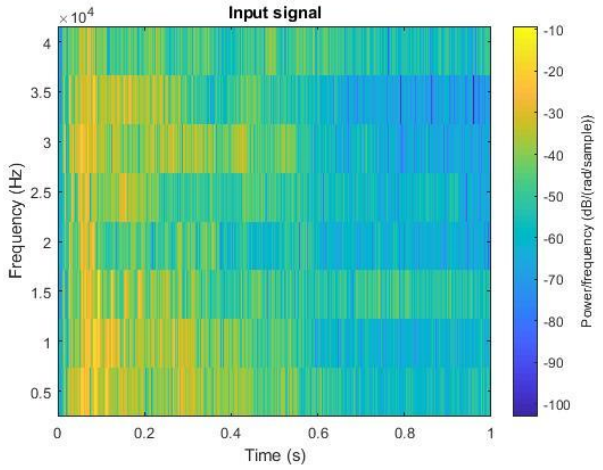


Figure 6.13: a comparative analysis of TIMIT input signal Mix1

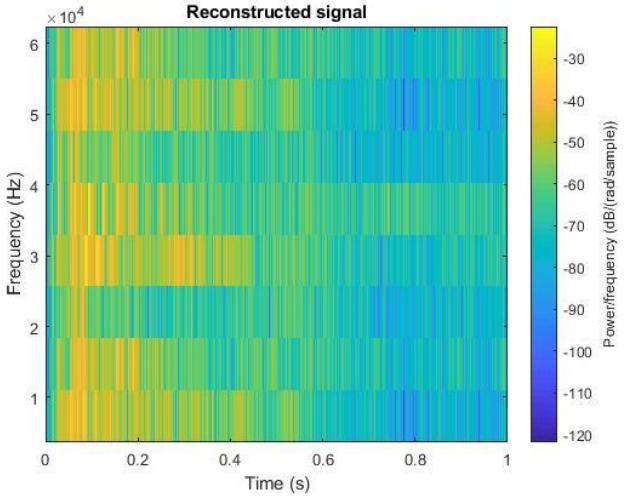


Figure 6.14: a comparative analysis of TIMIT reconstruction signal Mix2

The performance analysis of the input as well as the spectrogram reconstruction signals of mixtures 1 and 2 are shown in figures 6.13 and 6.14 above. When analysing the overhead statistics, our suggested approach produces cutting-edge results.

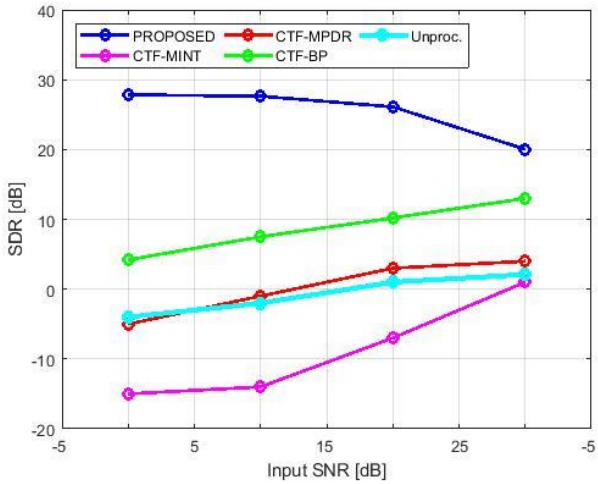


Figure 6.15: a comparative analysis of TIMIT input signal SDR based SNR

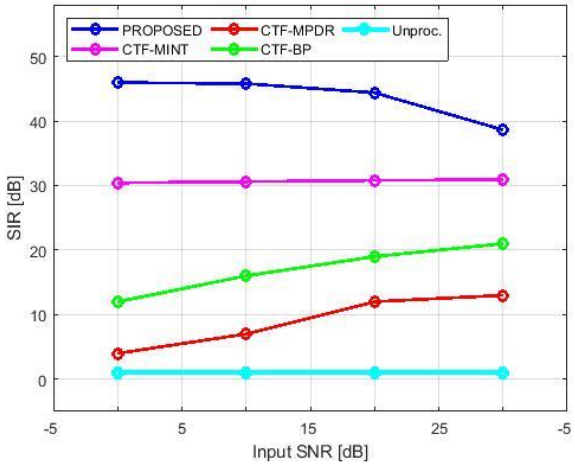


Figure 6.16: a comparative analysis of TIMIT input signal SIR based SNR

Figures 6.15 and 6.16 illustrate the analysis' findings as a function of the input signal-to-noise ratio from a combination of 4 microphones and 3 signal sources. There are two configurations for noise, i.e., 10-5 and 10-1, as was already mentioned. Our concept delivers better outcomes when compared to other existing techniques, such as CTF-MINT, CTF-MPDR, BP, and Unproc.

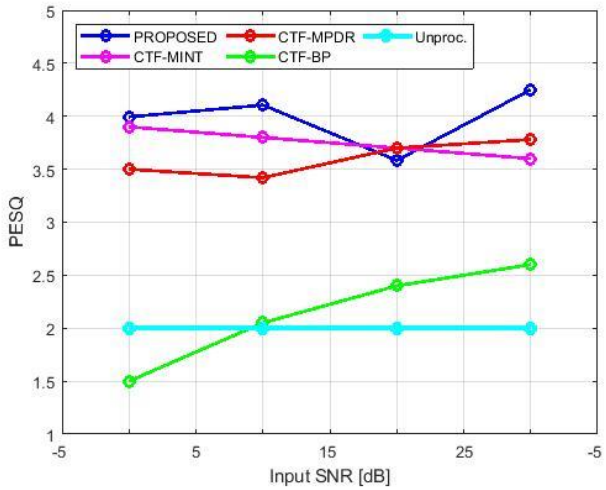


Figure 6.17: a comparative analysis of TIMIT reconstruction signal PESQ based SNR

Figure 6.17 displays the reconstructed signal based on the Perceptual Speech Quality Assessment's SNR spectrogram (PESQ). Here, the reverberation features are carefully assessed using the perceptual evaluation based on the PESQ SNR measurement. PESQ should be calculated for various sources while removing noise. When compared to figure 6.17, our suggested solution produces better results. Our suggested approach is contrasted with the other existing CTF-MINT, CTF-MPDR, BP, and Unproc strategies in this.

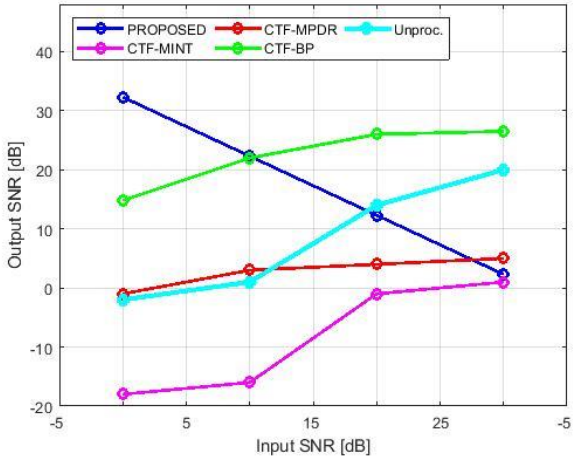


Figure 6.18: a comparative analysis of TIMIT reconstruction signal SNR

The performance analysis of the signal generated from the spectrogram reconstruction is shown in figure 6.18 above. The noise is being amplified if the input signal-to-noise ratio is more than 5 dB. Effective noise suppression requires that the signal-to-noise ratio at the output end always be higher than the signal-to-noise ratio at the input end. Our approach is contrasted with the current CTF-MINT, CTF-MPDR, BP, and Unproc techniques in this case.

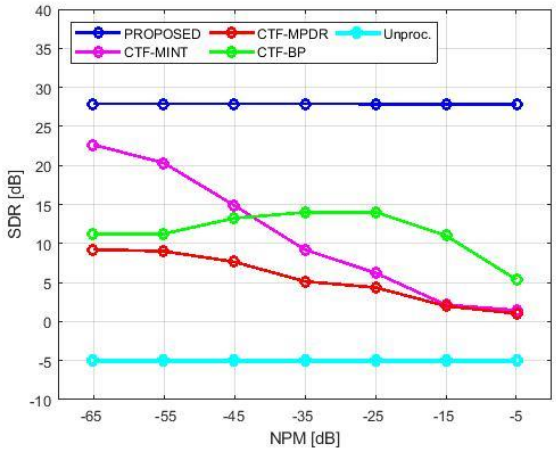


Figure 6.19: a comparative analysis of TIMIT SDR based NPM

Figure 6.19 depicts the correlation between the NPM and the mixing outcomes of 4 and 3 microphones. Both of the delta settings have been examined, just like in the prior experiment. Only the SDR indicator was examined. Our suggested strategy performs better than the other CTF-MINT, CTF-MPDR, BP, and Unproc existing techniques.

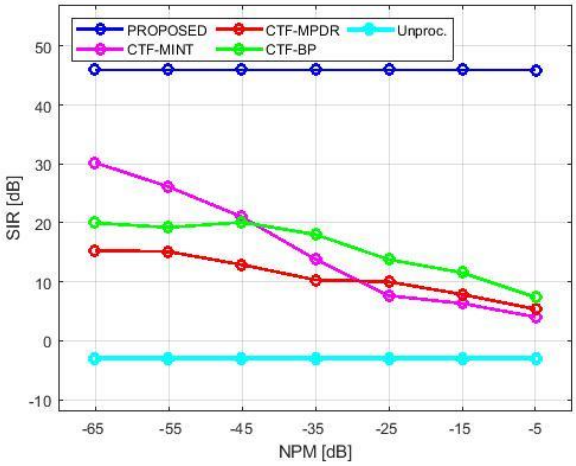


Figure 6.20: a comparative analysis of TIMIT SIR based NPM

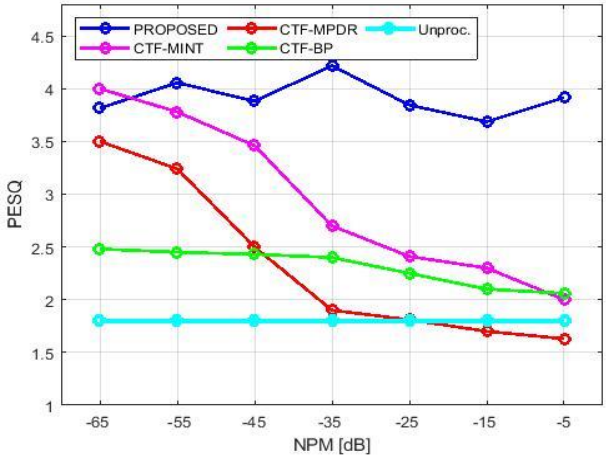


Figure 6.21: a comparative analysis of TIMIT PESQ based NPM

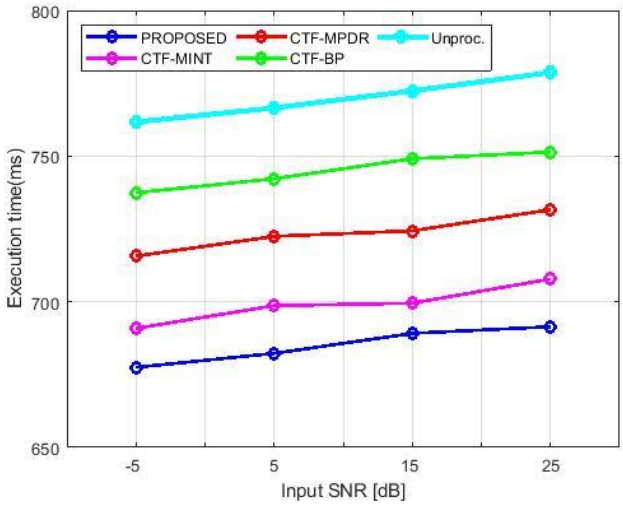


Figure 6.22: Comparative analysis of execution time

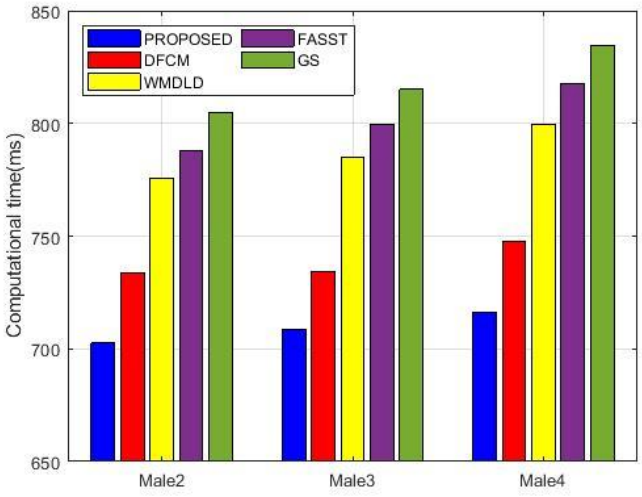


Figure 6.23: Comparative analysis of computational time

Figures 6.22 and 6.23 above show a study of execution and computational times side by side. The overall length of time that the process spends operating is known as the execution time; this time is typically independent of the commencement time but frequently depends on the input data. We frequently set deadlines for ongoing procedures, but we could also want to set one for a one-off process. Calculation time is the amount of time needed to complete a computation. The calculation time is inversely correlated with the number of rule applications when a computation is represented as a sequence of rule applications. Here, the current DFCM, WMDLD, FASST, and GS approaches are contrasted with our methodology.

6.9 Analysis of proposed methods in comparison to related work citations

Table 6.5 compares the suggested approaches with citations to similar works and lists the test's SDR, SIR, and SAR values for various signals. The dataset SiSEC2018 was initially used in comparative analysis to compare the experimental outcomes of the proposed system with those of the control group.

Table 6.6 compares the proposed methods with references to related literature, as well as the corresponding SDR and LLR values for the test on various signals. The second dataset from TIMIT compares the experimental outcomes of the proposed system. Tables 6.5 and 6.6 above demonstrate the comparison of the SiSEC2018 and TIMIT datasets. The suggested approach can produce better outcomes than the existing effort

Table 6.5: SiSEC2018 comparative analysis

Related work	**Year**		**Methodology**	**Outcome**		
				SDR	**SIR**	**SAR**
[36]	2020		DFCM	25.33	25.26	6.720
[37]	2019		MSS	25.53	25.16	6.302
[38]	2022		DGSS	15.43	25.46	6.223
[39]	2021		PSA	25.66	15.55	6.343
[40]	2021		ICASSP	25.44	15.44	8.552
[41]	2021		SESS	35.67	25.78	9.372
Proposed	-		**SCNN**	**45.53**	**35.16**	**10.372**

Table 6.6: comparative analysis of TIMIT

Related work	**Year**	**Methodology**	**Outcome**				
			SDR .diff	**SIR.diff**	**CDF**	**FWSeg.SNR**	**LLR**
[42]	2019	TFMM	6.6	4.22	-0.34	1.42	-0.11
[43]	2020	CSS	6.5	5.32	-0.44	1.42	-0.13
[44]	2020	SS	7.8	6.62	-0.34	1.52	-0.16
[45]	2021	SSDL	7.1	7.22	-0.24	1.72	-0.13
Proposed	**-**	**SCNN**	**8.1**	**10.22**	**-0.54**	**2.32**	**-0.19**

6.10. summary:

Multichannel speech separation is one of the most difficult challenges at the moment. KHMF and score-based CNN are combined and used for data sets (TIMIT, SiSEC-2011) to obtain results for multichannel source separation. The investigations findings demonstrate better performance when compared with the existing results. It has been observed that the results (performance evaluation metrics) are improved compared with existing works as indicated in the graphs and tables. The proposed SCNN method is calculated some performance measure which as SDR, SIR, and SAR. The value of the parameters is 45.53, 5.16 and 10.372 respectively. Experimental results show that our proposed approach accomplishes the most extreme SDR diff outcome of -5dB of 8.1. Comparable to the CTF-MINT, which achieves 8.05. The CTF-MPDR and CTF-BP had the SDR diff worst 7.71 and 7.4. The Unproc had the very worst SDR diff 5.71.

Chapter 7

Overall Conclusion and Future Scope of The Work

This chapter outlines the contributions of proposed research work for single channel and multi-channel source separation. In addition to the strength of research work that has been carried out, few limitations have also been observed which may be taken up as future scope of research.

7.1 Research findings of the thesis

The conclusions those have been discussed for each contribution are summarized as below:

In the method of TFNMF for single channel source separation, the data consists of two or more than the clean speech signals. The TFNMF integrated with SNDNN technique has been applied and results have been obtained. Experiments show that our proposed method achieves the highest gains in PESQ, STIO, SIR and SDR outcomes of 3.58, 0.7, 42 and 7.5 at -9 dB. It has been observed that the results (performance evaluation metrics) are improved compared with existing works as indicated in the graphs and tables.

In the method of IFRO for single channel source separation, the data consists of nosy speech signals. The integral fox ride optimization (IFRO) integrated with retrieval-based deep neural network technique has been applied and results have been obtained. Experiments show that our proposed method achieves the highest gains in SDR, SIR, SAR STIO, and PESQ outcomes of 10.9, 15.3, 10.8, 0.08, and 0.58, respectively. The Joint-DNN-SNMF obtains 9.6, 13.4, 10.4, 0.07, and 0.50, comparable to the Joint-DNN-SNMF. It has been observed that the results (performance evaluation metrics) are improved compared with existing works as indicated in the graphs and tables.

In the method of GOMF and EDNN for Multichannel source separation, Grasshopper Optimization-based Matrix Factorization (GOMF) and Enthalpy-based DNN are combined and used for data sets (SASSEC07, SiSEC-2010) to obtain results for multichannel source separation. Experimental results show that our proposed approach accomplishes the most extreme SNR outcome of − 6dB of 24.0523. Comparable to the DNN-JAT, which achieves 18.50032. The RNN and NMF-DNN had the worst SNR 13.45434 and 12.29991. The investigations findings demonstrate better performance when compared with the existing results. It has been observed that the results (performance evaluation metrics) are improved compared with existing works as indicated in the graphs and tables.

In the method of KHNF and SCNN for Multichannel source separation, KHMF and score-based CNN are combined and used for data sets (TIMIT, SiSEC-2011) to obtain results for multichannel source separation. Experimental results show that our proposed approach accomplishes the most extreme SDR dif outcome of − 5 dB of 8.1. Comparable to the CTF-MINT, which achieves 8.05. The CTF-MPDR and CTF-BP had the SDR dif worst 7.71 and 7.4. The Unproc had the very worst SDR dif 5.71. The investigations findings demonstrate better performance when compared with the existing results. It has been observed that the results (performance evaluation metrics) are improved compared with existing works as indicated in the graphs and tables.

7.2 Future scope

In the case of multi-channel source separation, sources may be mixed up with noise (Stationary and non-stationary) and investigations may be done using suitable methods. Supervised data sets have been considered in this work, but same methods may be extended for unsupervised data sets also

Work may also be extended for all other varieties of audio sources such as musical instruments sound sources mixed up with reverberations.

References

1. Daniel Patrick Whittlesey Ellis. Prediction-driven computational auditory scene analysis. PhD thesis, Massachusetts Institute of Technology, 1996.
2. Ray Meddis and Michael J Hewitt. Virtual pitch and phase sensitivity of a computer model of the auditory periphery. i: Pitch identification. The Journal of the Acoustical Society of America, 89(6):2866–2882, 1991.
3. Michael L Seltzer, Jasha Droppo, and Alex Acero. A harmonic-model-based front end for robust speech recognition. In Eighth European Conference on Speech Communication and Technology, 2003.
4. DeLiang Wang and Guy J Brown. Computational auditory scene analysis: Principles, algorithms, and applications. Wiley-IEEE press, 2006.
5. Guoning Hu and DeLiang Wang. Auditory segmentation based on onset and offset analysis. IEEE Transactions on Audio, Speech, and Language Processing, 15(2):396–405, 2007.
6. Patrik O Hoyer. Non-negative matrix factorization with sparseness constraints. Journal of Machine Learning Research, 5:1457–1469, 2004.
7. Andrzej Cichocki, Rafal Zdunek, and Shun-ichi Amari. New algorithms for non-negative matrix factorization in applications to blind source separation. In IEEE International Conference on Acoustics, Speech and Signal Processing, volume 5, 2006.
8. Mikkel N Schmidt and Rasmus Kongsgaard Olsson. Single-channel speech separation using sparse non-negative matrix factorization. In Proceedings of INTERSPEECH. ISCA, 2006.
9. Paris Smaragdis. Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs. In International Conference on Independent Component Analysis and Signal Separation, pages 494–499. Springer, 2004.
10. Paris Smaragdis. Convolutive speech bases and their application to supervised speech separation. IEEE Transactions on Audio, Speech, and Language Processing, 15(1):1–12, 2007.
11. Tuomas Virtanen. Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. IEEE Transactions on Audio, Speech, and Language Processing, 15(3):1066–1074, 2007.
12. R Mitchell Parry and Irfan Essa. Incorporating phase information for source separation via spectrogram factorization. In IEEE International Conference on Acoustics, Speech and Signal Processing, volume 2, pages II–661, 2007.

13. Y. Wang, A. Narayanan, and D. L. Wang, "On training targets for supervised speech separation," IEEE/ACM Trans. Speech Audio Processing, vol. 22, no. 12, pp. 1849-1858, Dec. 2014.
14. Zhuo Chen, Yi Luo, and Nima Mesgarani. Deep attractor network for singlemicrophone speaker separation. In IEEE International Conference on Acoustics, Speech and Signal Processing, pages 246–250, 2017.
15. S.Amari,A.Hyvarinen,S.Lee,T.W.LeeandS.A.David,"BlindSignalSeparationand Independent Component Analysis", Neurocomputing, vol. 49, pp.1-5, 2002.
16. J.F.Cardoso,"SourceSeparationusingHigherOrderMoments", in Proceedings ICASSP, Glasgow, 1989, pp. 2109-2112.
17. J.F.Cardoso,"BlindSignalSeparation:StatisticalPrinciples", in Proceedings of the IEEE, vol. 86, 1998, pp. 2009-2025.
18. E.Oja,J.Karhunen,L.WangandR.Vigario,"PrincipalandIndependentComponentsin Neural Networks", in Proc. VII Italian Workshop on Neural Nets WIRN, Italy, 1995.
19. C.JuttenandA.Taleb,"SourceSeparation:from dusktilldawn", in Proc. 2nd Int. Workshop on Independent Component Analysis and Blind Source Separation (ICA2000), Helsinki, Finland, 2000, pp. 12-26.
20. M. Girolami, Advances in Independent Component Analysis, Springer-Verlag, 2000.
21. S. Roberts and R. Everson, Independent Component Analysis: Principles and Practice, Cambridge Univ. Press, 2001.
22. A. Hyvarinen, J. Karhunen, and E. Oja, Independent component analysis and blind source separation, John Wiley & Sons pp.20–60, 2001.
23. D. Wang, "On ideal binary mask as the computational goal of auditory scene analysis" in Speech Separation by Humans and Machines, P. Divenyi, Ed. Norwell, MA: Kluwer, 2005, pp. 181–197.
24. J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2016, pp. 31–35.
25. T. Mikolov, M. Karafiát, L. Burget, J. ˇCernock`y, and S. Khudanpur, "Recurrent neural network-based language model," in Eleventh annual conference of the international speech communication association, 2010.
26. Christensen, H., Barker, J., Ma, N., et al.: 'The chime corpus: a resource and a challenge for computational hearing in multisource environments'. Eleventh Annual Conf. of the Int. Speech Communication Association, Chiba, Japan, 2010.

27. Y.-X. Wang, Y.-J. Zhang, "Nonnegative matrix factorization: A comprehensive review", IEEE Trans. Knowledge and Data Engineering, vol. 25, no. 6, pp. 1336- 1353, 2013.
28. D. D. Lee, H. S. Seung, "Algorithms for nonnegative matrix factorization", Proc. Advances in Neural Information Processing Systems (NIPS), pp. 556-562, 2001.
29. E. Vincent, N. Bertin, R. Gribonval, F. Bimbot, "From blind to guided audio source separation", IEEE Signal Processing Magazine, vol. 31, no. 3, pp. 107-115, 2014.
30. John R Hershey, Zhuo Chen, Jonathan Le Roux, and Shinji Watanabe. Deep clustering: Discriminative embeddings for segmentation and separation. In IEEE International Conference on Acoustics, Speech and Signal Processing, pages 31–35. IEEE, 2016.
31. Yusuf Isik, Jonathan Le Roux, Zhuo Chen, Shinji Watanabe, and John R Hershey. Single-channel multi-speaker separation using deep clustering. In Proceedings of INTERSPEECH, pages 545–549, 2016.
32. Zhong-Qiu Wang, Jonathan Le Roux, and John R Hershey. Alternative objective functions for deep clustering. In IEEE International Conference on Acoustics, Speech and Signal Processing, pages 686–690, 2018.
33. Yi Luo, Zhuo Chen, and Nima Mesgarani. Speaker-independent speech separation with deep attractor network. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 26(4):787–796, 2018.
34. Dong Yu, Morten Kolbæk, Zheng-Hua Tan, and Jesper Jensen. Permutation invariant training of deep models for speaker-independent multi-talker speech separation. In IEEE International Conference on Acoustics, Speech and Signal Processing, pages 241–245, 2017.
35. Morten Kolbæk, Dong Yu, Zheng-Hua Tan, and Jesper Jensen. Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 25(10):1901–1913, 2017.
36. L. Benaroya and F. Bimbot, Wiener based source separation with HMM/GMM using a single sensor,in Proc. 4th International Symposium on Independent Component Analysis and Blind Signal Separation (ICA2003), Nara, Japan, Apr. 2003, pp. 957-961.
37. SiSEC 2018: Signal Separation Evaluation Campaign, https://sisec.inria.fr.” [Online]. Available: http://sisec.inria.fr/ 2018-professionally-produced-music-recordings/
38. J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, “DARPA TIMIT acoustic phonetic continuous speech corpus CDROM,” 1993

39. Naoya Takahashi, Sudarsanam Parthasaarathy, Nabarun Goswami, and Yuki Mitsufuji. Recursive speech separation for unknown number of speakers. In Proceedings of INTERSPEECH, pages 1348–1352, 2019.
40. E. Vincent, C. Févotte, L. Benaroya, and R. Gribonval, A tentative typology of audio source separation tasks,in Proc. 4th International Symposium on Independent Component Analysis and Blind Signal Separation (ICA2003), Nara, Japan, Apr. 2003, pp. 715-720.
41. Cichocki and S. I. Amari, Adaptive Blind Signal and Image Processing –Learning Algorithms and Applications, J. Wiley & Sons Ltd., 2003.
42. Michael Stark, Michael Wohlmayr, and Franz Pernkopf. Source-filter-based single-channel speech separation using pitch information. IEEE Transactions on Audio, Speech, and Language Processing, 19(2):242–255, 2011.
43. N. Mohammadiha, "Speech enhancement using nonnegative matrix factorization and hidden Markov models," Ph.D. dissertation, KTH -Royal Inst. of Technol., Stockholm, Sweden, 2013.
44. T. G. Kang, K. Kwon, J. W. Shin and N. S. Kim, "NMF-based Target Source Separation Using Deep Neural Network," in IEEE Signal Processing Letters, vol. 22, no. 2, pp. 229-233, Feb. 2015. doi: 10.1109/LSP.2014.2354456.
45. S. Nie, S. Liang, W. Liu, X. Zhang and J. Tao, "Deep Learning Based Speech Separation via NMF-Style Reconstructions," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 26, no. 11, pp. 2043-2055, Nov. 2018.
46. A. Narayanan and D. Wang, "Investigation of Speech Separation as a Front-End for Noise Robust Speech Recognition," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 22, no. 4, pp. 826-835, April 2014. doi: 10.1109/TASLP.2014.2305833.
47. Z. Wang and D. Wang, "Combining Spectral and Spatial Features for Deep Learning Based Blind Speaker Separation," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 27, no. 2, pp. 457-468, Feb. 2019.
48. Y. Luo, Z. Chen and N. Mesgarani, "Speaker-Independent Speech Separation With Deep Attractor Network," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 26, no. 4, pp. 787-796, April 2018.
49. N. N. An, N. Q. Thanh and Y. Liu, "Deep CNNs With Self-Attention for Speaker Identification," in IEEE Access, vol. 7, pp. 85327-85337, 2019. doi: 10.1109/ACCESS.2019.2917470.

50. Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal timefrequency magnitude masking for speech separation," IEEE/ACM Trans. Audio Speech Lang. Process., vol. 27, Aug. 2019.
51. H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, “Phasesensitive and recognition-boosted speech separation using deep recurrent neural networks,” in 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2015, pp. 708–712.
52. Y. Isik, J. L. Roux, Z. Chen, S. Watanabe, and J. R. Hershey, “Single-channel multi-speaker separation using deep clustering,” arXiv preprint arXiv:1607.02173, 2016.
53. Y. Isik, J. L. Roux, Z. Chen, S. Watanabe, and J. R. Hershey, “Single-channel multi-speaker separation using deep clustering,” arXiv preprint arXiv:1607.02173, 2016.
54. E. Nachmani, Y. Adi, and L. Wolf, “Voice separation with an unknown number of multiple speakers,” in International Conference on Machine Learning. PMLR, 2020, pp. 7164–7175.
55. Dovrat, S., Nachmani, E., & Wolf, L. (2021). Many-speakers single channel speech separation with optimal permutation training. *arXiv preprint arXiv:2104.08955*.
56. Saleem, N., & Khattak, M. I. (2020). Deep neural networks based binary classification for single channel speaker independent multi-talker speech separation. *Applied Acoustics*, *167*, 107385.
57. Yi Luo and Nima Mesgarani. Tasnet: time-domain audio separation network for real-time, single-channel speech separation. In IEEE International Conference on Acoustics, Speech and Signal Processing, pages 696–700, 2018.
58. Keisuke Kinoshita, Lukas Drude, Marc Delcroix, and Tomohiro Nakatani. Listening to each speaker one by one with recurrent selective hearing networks. In IEEE International Conference on Acoustics, Speech and Signal Processing, pages 5064–5068, 2018.
59. Richard Lyon. A computational model of binaural localization and separation. In IEEE International Conference on Acoustics, Speech and Signal Processing, volume 8, pages 1148–1151, 1983.
60. Wang, Z. Q., Le Roux, J., & Hershey, J. R. (2018, April). Multi-channel deep clustering: Discriminative spectral and spatial embeddings for speaker-independent speech separation. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 1-5). IEEE.
61. Chen, Z., Yoshioka, T., Xiao, X., Li, L., Seltzer, M. L., & Gong, Y. (2018, April). Efficient integration of fixed beamformers and speech separation networks for multi-

channel far-field speech separation. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5384-5388). IEEE.

62. Perotin, L., Serizel, R., Vincent, E., & Guérin, A. (2018, April). Multichannel speech separation with recurrent neural networks from high-order ambisonics recordings. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 36-40). IEEE.
63. Wang, Z. Q., & Wang, D. (2018, September). Integrating Spectral and Spatial Features for Multi-Channel Speaker Separation. In *Interspeech* (pp. 2718-2722).
64. Gu, R., Wu, J., Zhang, S. X., Chen, L., Xu, Y., Yu, M., ... & Yu, D. (2019). End-to-end multi-channel speech separation. *arXiv preprint arXiv:1905.06286*.
65. Luo, Y., Chen, Z., Mesgarani, N., & Yoshioka, T. (2020, May). End-to-end microphone permutation and number invariant multi-channel speech separation. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6394-6398). IEEE.
66. Gu, R., Zhang, S. X., Chen, L., Xu, Y., Yu, M., Su, D., ... & Yu, D. (2020, May). Enhancing end-to-end multi-channel speech separation via spatial feature learning. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 7319-7323). IEEE.
67. Gu, R., Zhang, S. X., Zou, Y., & Yu, D. (2021). Complex neural spatial filter: Enhancing multi-channel target speech separation in complex domain. *IEEE Signal Processing Letters*, *28*, 1370-1374.
68. Zhang, Z., Xu, Y., Yu, M., Zhang, S. X., Chen, L., Williamson, D. S., & Yu, D. (2021). Multi-channel multi-frame ADL-MVDR for target speech separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *29*, 3526-3540.
69. Li, G., Yu, J., Deng, J., Liu, X., & Meng, H. (2022, May). Audio-visual multi-channel speech separation, dereverberation and recognition. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6042-6046). IEEE.
70. Kevin T Hill and Lee M Miller. Auditory attentional control and selection during cocktail party listening. Cerebral cortex, 20(3):583–590, 2009.
71. Neri, J., Badeau, R., & Depalle, P. (2021, August). Unsupervised blind source separation with variational auto-encoders. In *2021 29th European Signal Processing Conference (EUSIPCO)* (pp. 311-315). IEEE.
72. Drude, L., Hasenklever, D., & Haeb-Umbach, R. (2019, May). Unsupervised training of a deep clustering model for multichannel blind source separation. In *ICASSP 2019-*

2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 695-699). IEEE.

73. Drude, L., & Haeb-Umbach, R. (2019). Integration of neural networks and probabilistic spatial models for acoustic blind source separation. *IEEE Journal of Selected Topics in Signal Processing*, *13*(4), 815-826.
74. Drude, L., von Neumann, T., & Haeb-Umbach, R. (2018, April). Deep attractor networks for speaker re-identification and blind source separation. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 11-15). IEEE.
75. He, P., She, T., Li, W., & Yuan, W. (2018). Single channel blind source separation on the instantaneous mixed signal of multiple dynamic sources. *Mechanical systems and signal processing*, *113*, 22-35.
76. Seetharaman, P., Wichern, G., Le Roux, J., & Pardo, B. (2019, May). Bootstrapping single-channel source separation via unsupervised spatial clustering on stereo mixtures. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 356-360). IEEE.
77. J. Benesty, M. M. Sondhi, Y. Huang, Springer Handbook of Speech Processing, USA, NY, New York:Springer, 2007.
78. A.Hyvarinen,"Survey on Independent Component Analysis", Neural Computing Surveys, vol. 1, pp. 94-128, 1999.
79. R.Vigario,V.Joursmaki,M.Hamalainen,R.Hari,andE.Oja,"Independent Component Analysis for identification of artifacts in magnetoencephalographic recordings", in Advances in Neural Information Processing Systems 10, 1998, pp. 229-235.
80. Yi Luo and Nima Mesgarani. Conv-tasnet: Surpassing ideal time-frequency magnitude masking for speech separation. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 27(8):1256–1266, 2019.
81. Hennequin, R.; Khlif, A.; Voituret, F.; Moussallam, M. Spleeter: A fast and efficient music source separation tool with pre-trained models. J. Open Source Softw. 2020, 5, 2154.
82. Chang, C.H.; Chung, S.H.; Manthiram, A. Ultra-lightweight PANiNF/MWCNT-functionalized separators with synergistic suppression of polysulfide migration for Li–S batteries with pure sulfur cathodes. J. Mater. Chem. A 2015, 3, 18829–18834
83. Stephan Getzmann, Julian Jasny, and Michael Falkenstein. Switching of auditory attention in "cocktail-party" listening: ERP evidence of cueing effects in younger and older adults. Brain and cognition, 111:1–12, 2017.

84. G. Wichern, J. Antognini, M. Flynn, L. R. Zhu, E. McQuinn, D. Crow, E. Manilow, and J. Le Roux, “WHAM!: Extending speech separation to noisy environments,” in Proc. Interspeech, 2019, pp. 1368–1372.
85. Shi, J.; Xu, J.; Fujita, Y.; Watanabe, S.; Xu, B. Speaker-Conditional Chain Model for Speech Separation and Extraction. arXiv 2020, arXiv:2006.14149.
86. Liu, Y.; Delfarah, M.; Wang, D. Deep CASA for talker-independent monaural speech separation. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 6354–6358.
87. Nguyen, V.N.; Sadeghi, M.; Ricci, E.; Alameda-Pineda, X. Deep Variational Generative Models for Audio-visual Speech Separation. arXiv 2020, arXiv:2008.07191.
88. Luo, Y.; Chen, Z.; Yoshioka, T. Dual-path rnn: Efficient long sequence modeling for time-domain single-channel speech separation. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–9 May 2020; pp. 46–50.
89. Wang, D.; Chen, J. Supervised Speech Separation Based on Deep Learning: An Overview. IEEE/ACM Trans. Audio Speech Lang. Process. 2018, 26, 1702–1726.
90. Luo, Y.; Mesgarani, N. Conv-tasnet: Surpassing ideal time—Frequency magnitude masking for speech separation. IEEE/ACM Trans. Audio Speech Lang. Process. 2019, 27, 1256–1266.
91. Shi, Z.; Liu, R.; Han, J. La furca: Iterative context-aware end-to-end monaural speech separation based on dual-path deep parallel inter-intra bi-lstm with attention. arXiv 2020, arXiv:2001.08998.
92. Han, C.; Luo, Y.; Li, C.; Zhou, T.; Kinoshita, K.; Watanabe, S.; Delcroix, M.; Erdogan, H.; Hershey, J.R.; Mesgarani, N.; et al. Continuous Speech Separation Using Speaker Inventory for Long Multi-talker Recording. arXiv 2020, arXiv:2012.09727.
93. Fan, C.; Tao, J.; Liu, B.; Yi, J.; Wen, Z.; Liu, X. Deep attention fusion feature for speech separation with end-to-end post-filter method. arXiv 2020, arXiv:2003.07544.
94. Yi Luo and Nima Mesgarani. Real-time single-channel dereverberation and separation with time-domain audio separation network. In Proceedings of INTERSPEECH, pages 342–346, 2018.
95. Hao, D.D.; Tran, S.T.; Chau, D.T. Speech Separation in the Frequency Domain with Autoencoder. J. Commun. 2020, 15, 841–848.

96. Felix Weninger, Jonathan Le Roux, John R Hershey, and Shinji Watanabe. Discriminative nmf and its application to single-channel source separation. In Fifteenth Annual Conference of the International Speech Communication Association, 2014.
97. P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Deep learning for monaural speech separation," in 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2014, pp. 1562–1566.
98. J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2016, pp. 31–35.
99. S. Makino, T.-W. Lee, H. Sawada, Blind Speech Separation, USA, NY, New York:Springer, 2007.
100. Cherry E. Colin. Some experiments on the recognition of speech, with one and with two ears. The Journal of The Acoustical Society of America, 25(5): 975–979, 1953.
101. Paul D O'grady and Barak A Pearlmutter. Discovering convolutive speech phones using sparseness and non-negativity. In International Conference on Independent Component Analysis and Signal Separation, pages 520–527. Springer, 2007.
102. Tuomas Virtanen. Speech recognition using factorial hidden Markov models for separation in the feature space. In Ninth International Conference on Spoken Language Processing, 2006.
103. Trausti Kristjansson, John Hershey, Peder Olsen, Steven Rennie, and Ramesh Gopinath. Super-human multi-talker speech recognition: The IBM 2006 speech separation challenge system. In Ninth International Conference on Spoken Language Processing, 2006.

Publications

International Journals – Published

1. Koteswararao, Yannam Vasantha, and C. B. Rama Rao. "Single channel source separation using time–frequency non-negative matrix factorization and sigmoid base normalization deep neural networks." *Multidimensional Systems and Signal Processing* (2022): 1-21.
2. Koteswararao, Yannam Vasantha, and C. B. Rama Rao. "An Efficient Optimal Reconstruction Based Speech Separation Based on Hybrid Deep Learning Technique." *Defence Science Journal* 72.3 (2022).
3. Koteswararao, Yannam Vasantha, and C. B. Rama Rao. "Multichannel speech separation using hybrid GOMF and enthalpy-based deep neural networks." *Multimedia Systems* 27.2 (2021): 271-286.
4. Koteswararao, Yannam Vasantha, and C. B. Rama Rao. "Multichannel KHMF for speech separation with enthalpy based DOA and score based CNN (SCNN)." *Evolving Systems* (2022): 1-18.