

Masters of Science in Electronics and Telecommunication Engineering



Improved Speech Enhancement Through Attention-Driven T-F Masking Strategies

by

Khadija Akter

ID: 19METE012P

This thesis is submitted in partial fulfillment of the requirement for the degree of
MASTER OF SCIENCE IN ELECTRONICS AND TELECOMMUNICATION
ENGINEERING

Department of Electronics and Telecommunication Engineering
Chittagong University of Engineering and Technology

Chattogram-4349, Bangladesh.

November, 2024

CERTIFICATION

The thesis titled **Improved Speech Enhancement Through Attention-Driven T-F Masking Strategies** submitted by **Khadija Akter**, Roll No. **19METE012P**, Session **2019-2020** has been accepted as satisfactory in partial fulfillment of the requirement for the degree of **Master of Science in Electronics and Telecommunication Engineering** on **November 2024**.

BOARD OF EXAMINERS

1. _____
Dr. Md. Azad Hossain Chairman (Supervisor)
Professor & Head
Department of Electronics and Telecommunication Engineering
Chittagong University of Engineering & Technology (CUET)
Chittagong-4349

2. _____
Dr. Md. Jahedul Islam Member
Professor
Department of Electronics and Telecommunication Engineering
Chittagong University of Engineering & Technology (CUET)
Chittagong-4349

3. _____
Dr. Md. Saiful Islam Member
Associate Professor
Department of Electronics and Telecommunication Engineering
Chittagong University of Engineering & Technology (CUET)
Chittagong-4349

4. _____
Dr. Md. Mahbub Hasan Member (External)
Professor
Department of Electrical and Electronic Engineering
Khulna University of Engineering & Technology (KUET)
Khulna-9203

Author Declaration for Electronic Submission of MRP

I hereby declare that I am the sole author of the MRP. This is a true copy of the MRP, including any final revisions, as accepted by my Examiners.

I hereby also declare that the work contained in this Thesis could be uploaded to the repository of the library, Chittagong University of Engineering and Technology after 1 year of submission.

I authorize Chittagong University of Engineering and Technology to land this MRP to other institutions or individuals for the purpose of scholarly research.

I further authorize the Chittagong University of Engineering and Technology to reproduce this MRP by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

I understand that my MRP may be made electronically available to the public.

Signature of the Candidate

Khadija Akter
ID: 19METE012P

CANDIDATE'S DECLARATION

It is hereby declared that this thesis or any part of it has not been submitted elsewhere for the award of any degree or diploma.

Signature of the Candidate

Khadija Akter
ID: 19METE012P

APPROVAL by the SUPERVISOR

This is to certify that KHADIJA AKTER has carried out this research work under my supervision and that she has fulfilled the relevant Academic Ordinance of the Chittagong University of Engineering and Technology so that she is qualified to submit the following Thesis in the application for the degree of MASTER of SCIENCE in ELECTRONICS AND TELECOMMUNICATION ENGINEERING. Furthermore, the Thesis compiles with the PLAGIARISM and ACADEMIC INTEGRITY regulations of CUET.

Signature of the Supervisor

Dr. Md. Azad Hossain
Professor & Head
Department of Electronics and Telecommunication Engineering
Chittagong University of Engineering and Technology
Chittagong-4349

List of Publications

The following publication is a direct consequence of the research carried out during the elaboration of the thesis and gives an idea of the progression that has been achieved.

Journal Articles

1. Namey, A. A., **Akter, K.**, Hossain, M. A., & Dewan, M. A. A. (2024). CochleaSpecNet: An attention based dual branch hybrid CNN-GRU network for speech emotion recognition using cochleagram and spectrogram. *IEEE Access*. (Accepted for publication).
2. Orin, A. J., **Akter, K.**, Hossain, M. A., & Dewan, M. A. A. (2024). AD-FCINet: A fusion of convolution-involution network for Alzheimer's disease staging with channel and spatial-specific features. *IEEE Transactions on Medical Imaging*. (Submitted for publication).

Conference Proceedings

- **Akter, K.**, Mamun, N., Hossain, M. A. (2023, February). A tf masking based monaural speech enhancement using u-net architecture. In 2023 International Conference on Electrical, Computer and Communication Engineering (ECCE) (pp. 1-5). IEEE.

Dedication

To the storms that sought to break me, yet only taught me how to bend. The resistance and opposition fueled my perseverance and made this journey more meaningful.

Acknowledgment

I am profoundly thankful to Allah for granting me strength, direction, and blessings during my academic journey. I am wholeheartedly grateful to my supervisor, **Dr. Md. Azad Hossain**, Professor, and Head of the ETE Department at CUET, for his essential mentorship, consistent support, and expert assistance. His support has significantly motivated my academic success. I would also like to extend my heartfelt thanks to the members of my oral examination committee for their useful feedback and constructive suggestions. I want to express my sincere gratitude to the teachers and other ETE Department members for their knowledge-sharing, encouragement, and dedication to creating a dynamic learning environment.

Lastly, I would want to express my gratitude to my family and loved ones for their unwavering support, which have helped me work on my proposal with confidence throughout my life and not just in the present.

Abstract

Speech enhancement (SE) aims to elevate the perceptual quality and intelligibility of speech signals by mitigating ambient noise and distortions. Recent speech enhancement methods struggle with preserving fine temporal details in noisy environments, where low-resolution speech increases noise sensitivity. The difficulty of conveying these details, crucial for understanding speech, is compounded by the computational complexity of mapping-based approaches, which require learning the full range of clean spectrogram values. However, this work proposes a masking-based strategy to address these challenges in monaural speech. Two distinct approaches are explored by employing the Ideal Ratio Mask (IRM) into deep learning frameworks: a U-Net inspired architecture and a Time-Frequency Transformer Network (TF-TransNet). The effectiveness of both approaches is analyzed across different Signal-to-Noise Ratios (SNRs) using several objective metrics: Short-Time Objective Intelligibility (STOI), Perceptual Evaluation of Speech Quality (PESQ), Segmental Signal-to-Noise Ratio (SSNR) and Scale-Invariant Signal-to-Distortion Ratio (SI-SDR). In the first approach, the U-Net model comprises an encoder for feature extraction, a decoder for reconstructing the clean speech signal, and skip connections that enable the direct transfer of key information between the encoder and decoder. The U-Net approach demonstrates significant improvements in speech intelligibility and quality under moderate to high SNRs and familiar noise types, although its performance declines in low SNR conditions and with unseen noise types. To address these limitations, this work introduces the TF-TransNet, aiming to predict the IRM. This network utilizes a Time-Frequency Attention (TFA) encoder to process the noisy magnitude spectrogram, integrating the processed information with the decoder through a Time-Frequency (T-F) Transformer layer. By employing Multi-Head Self-Attention (MHSA) mechanisms and replacing the initial fully connected layer with a Long Short-Term Memory Unit (LSTM) in the T-F transformer layer, the model can dynamically prioritize informative features and their relationships across both time and frequency domains, enhancing its ability to capture long-range dependencies in sequential data. The proposed TF-TransNet outperforms existing models, such as the Convolutional Recurrent Network (CRN) and Gated Convolutional Recurrent Network (GCRN), demonstrating enhanced speech intelligibility and quality across both familiar and unfamiliar noise scenarios.

Keywords: Speech Enhancement, Masking, U-Net, Transformer, Speech Intelligibility, Speech Quality.

বিমূর্ত

স্পিচ এনহ্যান্সমেন্ট (SE) এর লক্ষ্য হল ব্যাকগ্রাউন্ডের শব্দ এবং বিকৃতি প্রশমিত করে বক্তৃতা সংকেতের উপলব্ধিগত গুণমান এবং বোধগম্যতাকে উন্নত করা। সাম্প্রতিক বক্তৃতা বর্ধিতকরণ পদ্ধতিগুলি কোলাহলপূর্ণ পরিবেশে সূক্ষ্ম অস্থায়ী বিবরণ সংরক্ষণের সাথে লড়াই করে, যেখানে কম-রেজোলিউশনের বক্তৃতা শব্দ সংবেদনশীলতা বাড়ায়। বক্তৃতা বোঝার জন্য অত্যন্ত গুরুত্বপূর্ণ এই বিবরণগুলি বোঝানোর অসুবিধা ম্যাপিং-ভিত্তিক পদ্ধতির কম্পিউটেশনাল জটিলতা দ্বারা জটিল হয়, যার জন্য পরিষ্কার বর্ণালীগ্রাম মানগুলির সম্পূর্ণ পরিসর শেখার প্রয়োজন হয়। তবে এই গবেষণায় একটি মাস্কিং-ভিত্তিক কৌশল প্রস্তাব করা হয়েছে, যা মনোরাল স্পিচের নয়েজ অবক্ষয় মোকাবেলা করে। আইডিয়াল রেশিও মাস্ক (IRM) ব্যবহার করে দুটি ভিন্ন পদ্ধতি গভীর শিক্ষা কাঠামোতে অনুসন্ধান করা হয়েছে: একটি ইউ-নেট অনুপ্রাণিত স্থাপনা এবং একটি টাইম-ফ্রিকোয়েন্সি ট্রান্সফরমার নেটওয়ার্ক (TF-TransNet)। উভয় পদ্ধতির কার্যকারিতা বিভিন্ন সিগন্যাল-টু-নয়েজ রেশিও (SNR) এর মধ্যে বিভিন্ন উদ্দেশ্যগত মেট্রিক ব্যবহার করে বিশ্লেষণ করা হয়েছে: শর্ট-টাইম অবজেকটিভ ইন্টেলিজিবিলিটি (STOI), পারসেপচুয়াল ইভালুয়েশন অফ স্পিচ কোয়ালিটি (PESQ), সেগমেন্টাল সিগন্যাল-টু-নয়েজ রেশিও (SSNR) এবং স্কেল-ইনভারিয়েন্ট সিগন্যাল-টু-ডিস্টোরশন রেশিও (SI-SDR)। প্রথম পদ্ধতিতে, (U-Net) মডেল একটি কাঠামো নিযুক্ত করে যা ইনপুট বক্তৃতা থেকে গুরুত্বপূর্ণ বৈশিষ্ট্যগুলি বের করার জন্য একটি এনকোডার, ক্লিন স্পিচ সিগন্যাল পুনর্গঠনের জন্য একটি ডিকোডার এবং এনকোডার এবং ডিকোডারের মধ্যে প্রয়োজনীয় তথ্যের সরাসরি স্থানান্তরকে সহজতর করতে স্কিপ সংযোগ অন্তর্ভুক্ত করে। (U-Net) পদ্ধতি মাঝারি থেকে উচ্চ SNR এবং পরিচিত শব্দের প্রকারের অধীনে বক্তৃতা বোধগম্যতা এবং গুণমানে উল্লেখযোগ্য উন্নতি দেখায়, যদিও এটির কর্মক্ষমতা কম (SNR) অবস্থাতে এবং অদেখা শব্দের প্রকারের সাথে হ্রাস পায়। এই সীমাবদ্ধতাগুলি মোকাবেলা করার জন্য, এই কাজটি টাইম-ফ্রিকোয়েন্সি ট্রান্সফরমার নেটওয়ার্ক (TF-TransNet) প্রবর্তন করে, যার লক্ষ্য আদর্শ অনুপাত মাস্ক (IRM) ভবিষ্যদ্বাণী করা। এই নেটওয়ার্কটি শব্দের ম্যাগনিটিউড স্পেকট্রোগ্রাম প্রক্রিয়া করতে একটি টাইম-ফ্রিকোয়েন্সি অ্যাটেনশন (TFA) এনকোডার ব্যবহার করে, প্রক্রিয়া করা তথ্যটি টাইম-ফ্রিকোয়েন্সি (T-F) ট্রান্সফরমার স্তরের মাধ্যমে ডিকোডারের সাথে সংহত করে। এই ইন্টিগ্রেশনটি কার্যকরভাবে স্পিচ সিগন্যালের অস্থায়ী এবং বর্ণালী বৈশিষ্ট্যগুলির মধ্যে জটিল নির্ভরতা ক্যাপচার করে এবং মডেল করে। মাল্টি-হেড সেলফ-অ্যাটেনশন (MHSA) মেকানিজম ব্যবহার করে এবং (T-F) ট্রান্সফরমার স্তরে প্রাথমিক সম্পূর্ণ সংযুক্ত স্তরটি লং শোর্ট-টার্ম মেমোরি ইউনিট (LSTM) দ্বারা প্রতিস্থাপন করে, মডেলটি সময় এবং ফ্রিকোয়েন্সি উভয় ডোমেইনে তথ্যপূর্ণ বৈশিষ্ট্য এবং তাদের সম্পর্কগুলি গতিশীলভাবে অগ্রাধিকার দেওয়ার সক্ষমতা অর্জন করে, যা ধারাবাহিক ডেটার দীর্ঘ-পরিসরের নির্ভরতাগুলি ধরতে সহায়তা করে। প্রস্তাবিত (TF-TransNet) বিদ্যমান মডেলগুলির তুলনায় উন্নত কার্যকারিতা প্রদর্শন করে, যেমন কনভোলিউশনাল রিকারেন্ট নেটওয়ার্ক (CRN) এবং গেটেড কনভোলিউশনাল রিকারেন্ট নেটওয়ার্ক (GCRN), পরিচিত এবং অচেনা শব্দের পরিস্থিতির উভয় ক্ষেত্রেই বক্তৃতার বোধগম্যতা এবং গুণমান বাড়ায়।

Table of Contents

Acknowledgment	i
Abstract	ii
বিমূর্ত	iii
1 Introduction	1
1.1 Research Background	1
1.2 Problem Statement	2
1.3 Motivation	2
1.4 Importance	3
1.5 Contributions	4
1.6 Thesis Layout	4
2 LITERATURE REVIEW	6
2.1 Monaural Speech Enhancement	6
2.2 Related Works	9
2.2.1 Conventional Statistical Signal Processing Methods	9
2.2.2 Deep Learning Based Speech Enhancemnet Methods	10
2.2.3 Learning Targets for Speech Enhancement Using Deep Learning	14
2.3 Research Gap	17
3 SE-UNet	18
3.1 Motivation	18
3.2 Proposed Methodology	18
3.2.1 Encoder-Decoder Architecture	19
3.2.2 Model Training and Optimization	22
3.3 Experimental Setup	22
3.3.1 Corpora Description	23
3.3.2 Tools and Libraries Used	23
3.4 Result Analysis	24
3.4.1 Objective Metric Evaluation	24
3.4.2 Spectrogram Analysis	25
3.4.3 Comparison with Mapping Strategies	26
3.5 Summary	26
4 TF-TransNet	27
4.1 Motivation	27
4.2 Proposed Methodology	27
4.2.1 Encoder Decoder Architecture	29
4.2.2 Time - Frequency Attention (TFA)	30

4.2.3	T-F Transformer Layer	32
4.3	Experimental Setup	35
4.3.1	Corpora Description	35
4.3.2	Tools and Libraries Used	36
4.4	Results	36
4.4.1	Evaluating Performance through Objective Metrics	36
4.4.2	Ablation Study of the Proposed Framework	37
4.4.3	Comparison with Existing Networks	38
4.4.4	Case Study	43
4.5	Summary	44
5	Conclusion	45
5.1	Discussion	45
5.2	Limitation	45
5.3	Future Scope	46
5.4	List of Publications	46

List of Figures

2.1	Monaural signal acquisition	6
3.1	Framework of the proposed T-F masking technique-based speech enhancement method	19
3.2	Architecture of the employed U-Net model for learning parameter optimization in the speech enhancement task, with distinct colors assigned to denote different operations in the encoder and decoder.	20
4.1	Block diagram illustrating the proposed TF-TransNet SE method, with spectral masking as its ceore technique.	28
4.2	Time-Frequency Attention mechanism employed in the Encoder. The input to the attention mechanism, $Input_{TFA}$ has dimensions (T , F , C), where T represents number of frames, F represents frequency bins, and C represents the number of channels.	31
4.3	Architecture of the TF-Transformer layer showcased in Fig. 4.1, consisting of a Time-Transformer and a Frequency-Transformer.Both maintain identical architectures.	33
4.4	Multihead attention mechanism that is employed in the bottleneck layer of the proposed architecture.	34
4.5	Spectrogram visualizations of (a) noisy and (b) clean speech samples, alongside the enhanced signal processed by existing SOTA models: (c) CRN, (d) GCRN, (e) Xiang et al. [50] , and proposed (f) TF-TransNet.	41
4.6	Spectrograms showing the difference between clean and enhanced, as well as noisy and enhanced speech signals at an SNR of 2.5 dB.	42
4.7	Unwrapped phase spectra of noisy, clean, and enhanced speech signals at an SNR of 2.5 dB, highlighting the TF-TransNet model's reliance on the noisy phase for reconstruction and the minimal phase discrepancies between clean and enhanced signals.	43
4.8	Spectrograms of the Noisy, Enhanced and Clean Speech Signals in a Practical Settings.	44

List of Tables

3.1	Layer-wise output dimensions	21
3.2	Optimum hyperparameters value used	22
3.3	Evaluation Under Seen and Unseen Conditions for Different Noise Types and SNR Values in the Suggested Approach	23
3.4	Performance of the Proposed Method in Terms of Speech Intelligibility and Quality	24
3.5	Comparison of Speech Enhancement Models Based on STOI, PESQ, SSNR Metrics	26
4.1	Seen and Unseen Conditions for Various Noise Types and SNR Values of the Proposed Work	36
4.2	Performance Comparison of Noisy and Proposed Methods Across Various Noise Types and SNR Levels	37
4.3	Ablation Study of the Proposed Network in Terms of Objective Intelligibility and Quality	38
4.4	Performance Comparison with Existing Models in Terms of Objective Intelligibility, Quality, and Distortion Scores for Three SNR Groups	39
4.5	Comparison with other SOTA models to Assess Objective Intelligibility, Quality, and Distortion Scores	39

Chapter 1: Introduction

1.1 Research Background

Speech enhancement is a field dedicated to refining and improving speech signals that have been compromised by noise, distortions, or other types of interference. The primary objective is to make these speech signals clearer and easier to understand for listeners or systems that rely on them. For instance, in environments with background noise—like a busy street, a crowded room, or even wind—the speech signal might become difficult to hear clearly. Speech enhancement techniques work to filter out this noise or correct distortions, so the spoken words become more distinct and intelligible.

This process is vital in many modern applications. For example:

1. Voice Assistants (e.g., Siri, Alexa): These systems need to accurately understand user commands, even in noisy environments. Speech enhancement helps improve the accuracy of voice recognition by ensuring the speech signal is as clear as possible.
2. Automatic Speech Recognition (ASR): This technology, used in things like transcription services or voice-controlled devices, relies on clean speech signals to function correctly. Enhancing the speech helps reduce errors in transcription or command recognition.
3. Hearing Aids: People with hearing impairments benefit from speech enhancement technologies in their hearing aids. These devices often have to amplify speech while minimizing background noise, making conversations easier to follow.
4. Mobile Communication: During phone calls, especially in noisy environments, speech enhancement improves call quality by reducing the background noise and enhancing the speaker's voice, ensuring both parties can communicate more effectively.

Real-time processing is a critical aspect of speech enhancement. It means that the enhancement occurs almost instantaneously as the speech is being received. This is essential for applications like live voice communication, where any delay can disrupt the conversation. The ability to process and enhance speech signals in real-time showcases the adaptability and importance of speech enhancement techniques across various industries and everyday scenarios.[1].

1.2 Problem Statement

Speech enhancement is sought after in many applications to improve either the perceived quality or the intelligibility of speech. Ideally, the goal is to enhance both aspects simultaneously. However, many acoustic enhancement approaches designed for background noise reduction prioritize quality improvement, frequently at the expense of adding distortions to the voice signal that may impair intelligibility. The challenge, and the primary objective of this thesis, is to eliminate or suppress background noise without distorting perceptions, with the primary aim of improving intelligibility and quality. Traditional approaches to this issue often presume a lack of correlation between speech and noise while considering noise patterns to be consistent. However, these methods frequently fall short when faced with dynamic and unpredictable real-world noise environments. In response to these limitations, recent research community has embraced deep neural networks (DNNs) as a promising alternative. The research investigates the application of attention-driven a transformer network to tackle this issue.

Also existing mapping techniques directly predict the clean speech signal from the noisy input, but they often struggle with generalization, particularly in complex or highly variable noise conditions. This can lead to artifacts and distortions in the enhanced speech, reducing its quality and intelligibility. In contrast, spectral masking has emerged as a more effective approach in many scenarios that selectively attenuates the noise components in the time-frequency domain while preserving the speech components. This approach is generally more robust, as it focuses on enhancing the speech signal by suppressing only the noise, rather than reconstructing the entire speech signal.

To address this problem,,this research aims to develop an attention-based spectral masking model using a time-frequency transformer to achieve superior noise suppression and speech enhancement compared to traditional mapping methods, especially in challenging monaural scenarios.

1.3 Motivation

In the realm of speech enhancement, two primary strategies have emerged: spectral mapping and spectral masking. Spectral mapping involves directly estimating the clean spectral magnitude from the noisy input, whereas spectral masking focuses on estimating a mask that selectively enhances or suppresses certain components of the speech signal. While spectral mapping has been a common approach, recent findings suggest that spectral masking offers a more efficient and effective solution.

By estimating a spectral mask, the model can concentrate on enhancing particular speech signal regions, leading to better noise suppression and preservation of speech quality. This is notably advantageous in complex noisy settings, where the distinction between speech and noise is subtle. In contrast, spectral mapping may struggle with these scenarios, as it requires the model to reconstruct the entire clean spectrum from noisy input, which can be more challenging and less precise. The superior performance of spectral masking over spectral mapping in enhancing speech signals motivates this research.

Over the past few years, the emergence of deep learning techniques has considerably advanced the field, with T-F masking standing out as an especially promising approach. T-F masking strategies involve estimating masks that can selectively enhance or suppress certain components of the speech signal in the time-frequency domain[2]. However, conventional T-F masking methods may still face limitations, especially in complex noise environments, where distinguishing between speech and noise components can be difficult.

1.4 Importance

The market for speech and audio processing applications is experiencing rapid growth, with smart speakers being a prime example. Valued at \$11.9 billion in 2019, the smart speaker market is projected to surpass \$35.5 billion by 2025, emphasizing the increasing demand for advanced speech processing technologies.

One of the key areas within this field is speech enhancement, which plays a crucial role in improving the clarity and quality of voice communication. As voice-activated devices, virtual assistants, and smart home systems become more integrated into daily life, ensuring accurate and reliable speech recognition is essential. Background noise, reverberations, and other distortions can significantly degrade the quality of audio input, affecting both user experience and the performance of voice-based systems.

Speech enhancement addresses these challenges by filtering out noise and improving the intelligibility of speech in real-world environments. This not only leads to better voice interactions but also enhances accessibility for users with hearing impairments, improves communication in noisy settings, and boosts the overall efficiency of voice-driven technologies. As the market continues to grow, the importance of developing high-performance speech enhancement algorithms becomes even more critical for ensuring seamless and effective communication across various applications.

Additionally, traditional speech enhancement methods, which rely on statistical models, often fail in real-world environments because of irrational pre-

sumptions, such as voice and noise being stationary or unrelated. On the other hand, DNN-based techniques have advanced significantly by approaching speech enhancement as a supervised learning problem.. These methods can handle complex, non-stationary noise without relying on prior assumptions, making them far more effective in real-world scenarios. In light of these benefits, this study suggests a T-F masking-based noise suppression method using a TF-TransNet to further improve performance in challenging acoustic conditions.

1.5 Contributions

The main contributions of this work are outlined as follows:

- This work introduces a masking-based strategy for speech enhancement using the IRM to mitigate noise degradation in monaural speech. This approach diverges from traditional mapping-based techniques by targeting the specific impact of noise on the speech spectrum.
- The thesis initially presents a U-Net-inspired design that effectively leverages skip connections between the encoder and decoder to facilitate information transfer, resulting in notable improvements in speech intelligibility and quality under moderate-to-high SNRs with familiar noise conditions.
- A distinct framework, TF-TransNet is proposed, equipped with a TFA encoder and a Time-Frequency Transformer layer. This model dynamically captures complex dependencies between temporal and spectral features, addressing the limitations of the U-Net model under challenging noise conditions.
- The effectiveness of both models is evaluated across various SNR levels using objective metrics such as STOI, PESQ, SSNR and SISDR.

1.6 Thesis Layout

The structure of the remaining thesis is as follows. Chapter 2 provides an extensive analysis of the literature, outlining important terms and providing an overview of the current research in the subject of SE task. Chapter 3 focuses on masking-based speech enhancement with U-Net framework, describing the proposed model architecture and its implementation for improving speech quality. Chapter 4 explores speech enhancement using Time-Frequency Transformer Network, outlining the methodology and performance analysis of the proposed

approach. Finally, Chapter 5 discusses the conclusions drawn from the findings and offers suggestions for future research directions

Chapter 2: LITERATURE REVIEW

This chapter will introduce and explain several important terms that are crucial for understanding the discussions in the upcoming sections. These key terminologies will lay the foundation for subsequent content. Following the explanation of these terms, a comprehensive review of the existing literature on speech enhancement will be presented.

2.1 Monaural Speech Enhancement

Audio signals typically come in two common formats: mono (one channel) and stereo (two channels). Monaural speech enhancement relies on audio captured from one microphone, making the task more challenging due to the absence of spatial information. These methods must differentiate speech from noise using only the data from a single source, which is particularly difficult in noisy or complex environments. In contrast, stereo methods utilize multiple microphones, capturing spatial information and employing techniques like beamforming to better separate speech from noise, offering improved accuracy but requiring more complex setups.

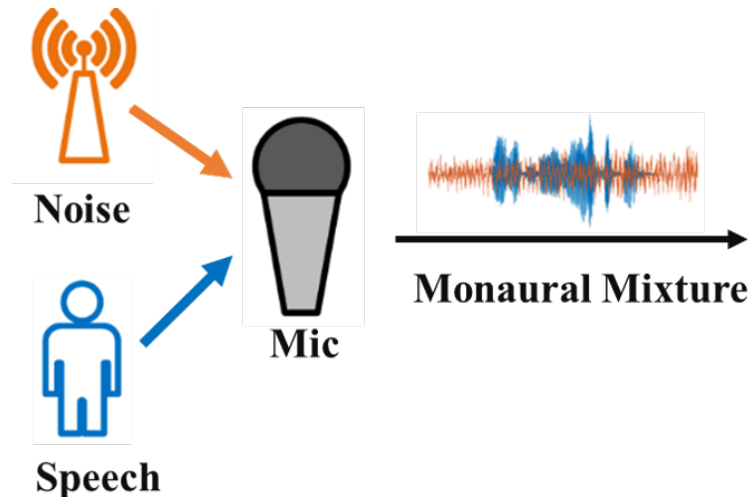


Figure 2.1: Monaural signal acquisition

When assessing the performance of SE systems, it is essential to use both subjective and objective measures. These measurements offer thorough information

about the improved speech’s quality and comprehensibility. Subjective evaluations involve human listeners who rate their perceptions of speech quality. Key subjective measures include:

- **Mean Opinion Score (MOS):** A higher score on a scale of 1 to 5 denotes a higher perceived quality of the improved speech, according to this metric.
- **Signal Distortion (SIG):** Also rated on a scale of 1 to 5, SIG reflects the clarity of the enhanced signal, with higher values suggesting reduced distortion.
- **Background Noise Intrusiveness (BAK):** This measure assesses how intrusive background noise is, using the same 1 to 5 scale, where a higher score denotes less perceptible noise.

In contrast to subjective measures, objective metrics provide quantifiable evaluations without human input. Important objective measures include:

- **STOI:** STOI is an objective metric that closely correlates speech intelligibility with human auditory perception. It begins by dividing both the clean (reference) and degraded (processed) signals into overlapping short-time frames. These frames are transformed into one-third octave bands using a filter bank to mimic the human auditory system’s frequency selectivity [3]. Subsequently, temporal envelopes are extracted from these bands using a low-pass filter. The metric computes the correlation coefficient between the corresponding clean and processed temporal envelopes for each band over a fixed-length frame. The final STOI score is the average of these correlation coefficients, normalized between 0 and 1. Higher scores indicate better alignment and higher intelligibility.
- **PESQ:** The PESQ metric assesses speech quality by comparing a reference signal to a degraded signal, addressing distortions such as time delays and spectral mismatches. Initially, PESQ applies a perceptual model to the signals, transforming them into perceptually weighted loudness spectra [4]. The model compensates for linear distortions, such as amplitude scaling and time alignment. Following this alignment, PESQ calculates the error by comparing the loudness spectra across frames. The distortion is weighted to account for human auditory masking effects. The final PESQ score, which ranges from -0.5 to 4.5, is derived from a combination of frame-based error metrics, with higher scores indicating better perceived quality.

- **SSNR:** The SSNR is another objective metric that can be utilized to evaluate speech enhancement tasks in either the time or frequency domain. It calculates the average ratio of short segments between the original and enhanced signals to derive the metric score[5]. The signal is segmented into fixed-length frames, typically 20–30 ms. For each segment, the SNR is computed as:

$$\text{SNR}(i) = 10 \log_{10} \left(\frac{\sum_n \text{Clean}(n)^2}{\sum_n (\text{Clean}(n) - \text{Enhanced}(n))^2} \right) \quad (2.1)$$

where $\text{Clean}(n)$ and $\text{Enhanced}(n)$ are the samples of the clean and enhanced signals, respectively, in frame i . Frames with extremely low energy may be excluded to avoid instability in the metric. The SSNR score is the average SNR across all valid frames. Higher SSNR values indicate greater noise suppression while preserving signal content.

- **SI-SDR:** The SI-SDR is an objective metric that quantifies the distortion between a reference signal and an estimated signal, focusing on the alignment of their content while being invariant to scaling differences. The optimal scaling factor α is computed as:

$$\alpha = \frac{\sum_t \text{reference}[t] \cdot \text{estimation}[t]}{\sum_t \text{reference}[t]^2} \quad (2.2)$$

Using α , the target signal is generated as:

$$e_{\text{target}} = \alpha \cdot \text{reference} \quad (2.3)$$

The SI-SDR score is then calculated as:

$$\text{SI-SDR} = 10 \log_{10} \left(\frac{\|e_{\text{target}}\|^2}{\|e_{\text{res}}\|^2} \right) \quad (2.4)$$

where $e_{\text{res}} = \text{estimation} - e_{\text{target}}$ is the residual noise. Higher SI-SDR values indicate less distortion, signifying a closer match between the estimated and reference signals.

This work primarily focuses on single-channel speech enhancement, recognizing its challenges but also its broad applicability. The chapter reviews progress in this area from two perspectives: traditional statistical signal processing methods, which rely on mathematical models to enhance speech, and cutting-edge deep learning approaches, which leverage neural networks and large datasets to

achieve state-of-the-art performance. The integration of these methods represents the evolution of the field, with deep learning showing particular promise in overcoming the limitations of traditional approaches by learning more complex patterns in speech and noise.

2.2 Related Works

2.2.1 Conventional Statistical Signal Processing Methods

Traditional speech enhancement techniques have long been grounded in statistical signal processing. These methods rely on mathematical models to estimate and suppress noise from speech signals. Commonly used approaches include spectral subtraction [6], subspace methods [7] and Wiener filtering [8] which were widely used due to their simplicity and well-established theoretical foundations. Spectral subtraction, for instance, assumes that noise can be modeled and subtracted from the noisy signal. However, it often encountered difficulties with non-stationary noise and introduced artifacts such as "musical noise" [6]. Another approach, subspace methods leveraged the idea that speech signals lie in a low-dimensional subspace, allowing for effective separation from noise through projections [7]. Although theoretically robust, subspace methods were computationally intensive and required accurate modeling of the subspaces, which might not always be feasible in practical applications. Furthermore, they often assumed a static noise model, limiting their effectiveness in dynamic environments. To address the limitations of spectral subtraction, Wiener filtering was emerged. This method offers a more refined approach by employing a gain function that acts as a linear filter on the noisy speech spectrum. In order to improve its efficacy in generating cleaner audio output, this gain function is constructed by reducing the mean square error (MSE) between the estimated and clear speech spectra.

Another statistical speech enhancement method is the Bayesian estimators that focus on minimizing the expected squared error between the estimated and actual clean signal. The above mentioned traditional speech enhancement methods are built upon assumptions that often do not align with real-world conditions. A common presumption is that speech and noise are uncorrelated, alongside the assumption that noise is stationary or at least more stationary than the speech signal. These assumptions can lead to significant inaccuracies in estimating the underlying model statistics, which is particularly problematic when dealing with highly non-stationary noises or unpredictable real-world environments. As a result, these methods frequently struggle to effectively suppress noise in challenging acoustic scenarios. In contrast, data-driven methods such as non-negative matrix

factorization (NMF) [9] avoid these simplifications but rely on other assumptions, such as sparse spectral representation, which can restrict their effectiveness.

2.2.2 Deep Learning Based Speech Enhancement Methods

There has been a noticeable move toward deep learning-based methods for speech improvement in order to overcome these constraints. Unlike traditional methods, deep learning models can learn complex patterns and dependencies directly from data, allowing them to adapt to varying noise characteristics and dynamic environments. These models leverage large datasets to develop robust representations of speech and noise, enabling more accurate separation and enhancement. Furthermore, deep learning techniques can incorporate contextual information, improving their ability to handle unexpected conditions that traditional methods might misinterpret. The shift to sophisticated deep learning techniques is a major advancement in addressing the drawbacks of traditional SE approaches, offering enhanced performance and reliability in diverse real-world applications.

Early approaches employed simpler models for speech enhancement task such as Multilayer Perceptrons (MLPs) [10, 11], Gaussian mixture models (GMMs) [12], and support vector machines (SVMs) [13]. Over time, focus shifted towards more advanced models like deep neural networks (DNNs) [14] [15] [16], which include at least one hidden layer, often implemented as multilayer perceptrons (MLPs). Karjol et al. [17] developed a speech enhancement method that utilizes multiple deep neural networks (DNNs). In their system, four DNNs, each three layers deep, collaborate to produce enhanced speech outputs, with a gating network assigning weights to combine these outputs. The approach achieved an average PESQ score of 2.65 for familiar noise and 2.19 for unfamiliar noise across a SNR range of -5 to 10 dB, evaluated on the TIMIT corpus. In study [18], a DNN masking technique reached a PESQ score of 2.705, demonstrating improved performance. Gao et al. [19] proposed a progressive learning framework for DNN-based SE to improve performance in low SNR conditions. Unlike conventional DNN methods that directly map noisy speech to clean speech, their approach decomposes the DNN into multiple stages, each designed to handle a specific SNR level. The model was trained using both single-SNR and multi-SNR configurations on the WSJ corpus and evaluated on a variety of noise types, including babble, factory, and destroyer engine noises. At SNRs of -5 and 5 dB, the suggested approach produced an average PESQ score of 1.93 for single-SNR training and 1.82 for multi-SNR training.

A denoising autoencoder (DAE) introduced by [20], is a neural network ar-

chitecture used for feature learning and noise reduction in data, making it particularly valuable for speech enhancement tasks. It is an extension of the basic autoencoder architecture, which encodes the input into a lower-dimensional space and subsequently decodes it back to its original form, with the aim of learning a compressed representation of the data. The key distinction of a DAE lies in its ability to handle noisy input data. During training, the input is intentionally corrupted by adding noise, and the model is trained to reconstruct the clean version of the input from the corrupted one. The fundamental operations of an autoencoder, including denoising autoencoders, are defined by the encoding and decoding processes. In the encoding stage, the input $\mathbf{x}$, which can be a noisy speech signal in the context of a denoising autoencoder, is mapped to a hidden representation $\mathbf{h}$ through a function:

$$\mathbf{h} = f_{\theta}(\mathbf{x}) = \sigma(\mathbf{W}_e \mathbf{x} + \mathbf{b}_e), \quad (2.5)$$

Here, $\mathbf{W}_e$ and $\mathbf{b}_e$ represent the encoder's weight matrix and bias vector, respectively, while σ denotes an activation function, such as ReLU or sigmoid. The hidden representation $\mathbf{h}$ is then transformed back into an approximation of the input $\hat{\mathbf{x}}$ during the decoding stage using the function:

$$\hat{\mathbf{x}} = g_{\theta'}(\mathbf{h}) = \sigma'(\mathbf{W}_d \mathbf{h} + \mathbf{b}_d), \quad (2.6)$$

where $\mathbf{W}_d$ and $\mathbf{b}_d$ are the decoder's weight matrix and bias vector, and σ' is the activation function in the decoding process. The output $\hat{\mathbf{x}}$ aims to reconstruct the clean version of the input $\mathbf{x}$. To train the autoencoder, a loss function, such as MSE, is used to minimize the difference between the clean input and the reconstructed output, expressed as:

$$\mathcal{L}(\mathbf{x}, \hat{\mathbf{x}}) = \|\mathbf{x} - \hat{\mathbf{x}}\|^2. \quad (2.7)$$

This training objective allows the denoising autoencoder to learn how to map noisy input data to clean outputs, thus effectively reducing noise in the processed signal.

The shift from DNN-based methods to Recurrent Neural Networks (RNNs), especially LSTM models, in the field of speech enhancement has been primarily driven by the limitations of traditional fully connected (FC) networks. While DNNs often consist of a large number of parameters, making them computationally intensive to train and deploy [21], they also struggle to effectively model the temporal characteristics inherent in speech signals. RNNs, on the other hand, offer a distinct advantage in this area due to their ability to model temporal

dependencies and maintain dynamic context over time, which is crucial for handling time-varying audio signals and noise environments [21], [22]. As a result, RNNs have become more suitable for tasks that involve varying-length speech data, providing a more efficient and context-aware approach to speech enhancement. Maas et al. [21] explored the use of RNNs for denoising input features such as MFCCs and found that RNNs outperformed DNNs in this context. Their study showed that RNNs were more effective in enhancing speech signals contaminated with noise, achieving better results across a range of SNR levels. This highlights the RNN’s ability to handle temporal dependencies in speech data, making it more suitable for improving speech quality under varying noise conditions compared to traditional DNN approaches. Another study highlighted the effectiveness of RNNs relative to DNN architectures in speech separation tasks [23]. LSTM networks were introduced to address the vanishing and exploding gradient issues commonly faced by RNNs in sequence learning, addressed by Jitong et al [24] for speech separation. Additionally, a modified version of LSTM, known as bidirectional LSTM (BLSTM), was investigated in the task of automatic speech recognition task[25]. This architecture processes information from both past and future states, enabling it to leverage the full contextual information of the input signal for better output sequence calculations.

While RNNs are adept at modeling temporal sequences, they struggle with effectively extracting local spatial features from spectrograms, which are time-frequency representations of speech conveying phonetic and acoustic information. RNNs did not inherently possess mechanisms to capture these local patterns within the spectrogram, making them less effective in scenarios where these patterns are significant. Convolutional Neural Networks (CNNs) became more relevant in speech processing as researchers sought models that could better extract spatial features from spectrograms [26], [27]. By applying convolutional layers, CNNs are able to capture localized patterns, such as energy concentrations across frequency bands or short-term temporal changes, which are critical for identifying phonetic elements and noise characteristics.

The integration of LSTM layers with Convolutional Neural Networks (CNNs) to form Convolutional Recurrent Networks (CRNs) represents a significant advancement in the processing of complex data types, particularly in speech and video analysis [28]. In this framework, the CNN component serves as an encoder, extracting spatial features from input data, such as spectrograms in speech processing, where it identifies important local patterns and phonetic elements. Following the CNN encoder, LSTM layers are incorporated to effectively model the temporal dependencies of the extracted features. LSTMs are designed to maintain a form of memory that captures long-range dependencies, which is crucial in

tasks where the context of previous frames influences the interpretation of current data. The combination of these two architectures enhances feature extraction and improves temporal modeling, allowing CRNs to adapt to varying noise conditions and providing a richer representation of the input data. This approach is particularly beneficial in applications such as speech recognition, where accurate understanding of both local acoustic features and long-term dependencies significantly enhances performance.

The integration of attention mechanisms in speech enhancement has been adopted in speech enhancement[29], motivated from their demonstrated effectiveness in various sequence-to-sequence learning tasks, such as machine translation, speech recognition, and keyword spotting [30] [31] [32] [33]. These attention-based models capitalize on the ability to dynamically prioritize certain aspects of input data, which aligns well with human auditory perception. In natural listening environments, humans can focus their attention on specific regions of an audio stream, such as the target speech, while simultaneously processing surrounding sounds, like background noise or interference, with less emphasis. This selective focus allows listeners to adjust their attention over time based on the relevance and importance of different audio components.

Transformer neural networks [34] have recently achieved state-of-the-art performance across various NLP tasks. Unlike RNNs, Transformers can process input sequences in parallel, enabling more efficient computation. Inspired from the effectiveness of transformer in NLP task Wei et al.[35] introduced a novel sequence learning model called SETransformer for the task of speech enhancement. The SETransformer utilizes a multi-layer architecture that combines local LSTMs, a transformer equipped with a multi-head attention mechanism, and a 1D convolutional network. Another approach introduces an enhanced transformer that enables direct context-aware modeling of speech sequences, resulting in improved speech separation performance [36]. To address the need for sequence order information without relying on positional encodings, it integrates a recurrent neural network into the Transformer architecture. Additionally, this enhanced transformer is embedded into a dual-path network, making the method effective for modeling extremely long speech sequences.

The two primary categories of deep learning techniques for monaural speech enhancement are: T-F domain methods and end-to-end time-domain techniques. T-F domain methods rely on the STFT to convert audio signals into time and frequency representations. This allows models to work on smaller, more manageable time-frequency components, and then reconstruct the enhanced audio using the inverse STFT. These methods are advantageous in that they exploit well-established signal processing techniques, but they may introduce artifacts during

the transformation and reconstruction processes. End-to-end time-domain methods, on the other hand, operate directly on raw audio waveforms. These methods bypass the need for intermediate transformations, enabling deep learning models to learn how to directly separate and enhance audio sources from the time-domain signal itself. This can result in a more natural sound quality and is particularly effective for cases where STFT may struggle with the complexity of the signal.

2.2.3 Learning Targets for Speech Enhancement Using Deep Learning

Within these approaches, deep learning models for speech enhancement typically use one of two training objectives: masking-based or mapping-based targets. Masking-based methods aim to estimate a mask that, when applied to the noisy input, isolates the clean speech signal. Mapping-based methods, however, focus on directly predicting the clean speech signal from the noisy input. Both strategies have their advantages, with masking-based approaches often being more interpretable and mapping-based approaches sometimes offering better performance in complex scenarios.

2.2.3.1 Mapping Strategies

In mapping-based speech enhancement (SE), the goal is to train a deep learning model to directly predict the clean speech signal, $\hat{s}(t)$, from a given noisy input, $x(t)$. The noisy input signal is typically represented as

$$x(t) = s(t) + n(t),$$

where $s(t)$ is the clean speech signal, and $n(t)$ represents the noise. A deep learning model f_θ , parameterized by θ , is then used to estimate the clean speech from the noisy input, resulting in the prediction

$$\hat{s}(t) = f_\theta(x(t)).$$

To train the model, a loss function L is defined, which quantifies the difference between the predicted clean speech $\hat{s}(t)$ and the ground truth clean speech $s(t)$. This is often done using mean squared error (MSE) or similar metrics, expressed as

$$L(\theta) = \frac{1}{T} \sum_{t=1}^T \|s(t) - \hat{s}(t)\|^2,$$

where T is the total number of time steps. The training process aims to minimize this loss function, enabling the model to learn a direct mapping from noisy speech to clean speech.

2.2.3.2 Masking Strategies

The masking-based approach for SE involves creating masks that represent the relationship between clean speech and noise in the time-frequency domain of a noisy audio signal. These masks help to isolate the clean speech components from the noise. Various masking-based targets have been developed, each aiming to improve different aspects of SE performance:

1. Ideal Binary Mask (IBM): The IBM assigns a binary value to each time-frequency unit in the spectrogram based on a local SNR threshold. It is defined as:

$$\text{IBM}(t, f) = \begin{cases} 1, & \text{if } \text{SNR}(t, f) > \tau \\ 0, & \text{otherwise} \end{cases} \quad (2.8)$$

where $\text{SNR}(t, f)$ is the SNR at time t and frequency f , and τ is a threshold value. IBM is known for improving speech intelligibility but results in sharp, binary labeling.

2. Ideal Ratio Mask (IRM): To address the limitations of the IBM due to its sharp binary labeling, the IRM[37] was introduced to provide a more refined mask, using real-valued ratios that compare the power of the clean speech $|C(t, f)|^2$ to that of the noisy signal $|N(t, f)|^2$:

$$\text{IRM}(t, f) = \sqrt{\frac{|C(t, f)|^2}{|C(t, f)|^2 + |N(t, f)|^2}} \quad (2.9)$$

where $|C(t, f)|$ and $|N(t, f)|$ are the magnitudes of the clean speech and noise at time t and frequency f .

3. Spectral Magnitude Mask (SMM): This mask focuses on enhancing the spectral magnitude of the clean speech while suppressing noise. It estimates the magnitude spectrum of clean speech:

$$\text{SMM}(t, f) = \frac{|C(t, f)|}{|N(t, f)|} \quad (2.10)$$

This approach helps in isolating the desired speech signal from noise by emphasizing the spectral strength of clean speech.

4. Phase-Sensitive Mask (PSM): While earlier methods primarily focused on magnitude, PSM incorporates phase information. It takes into account the

phase difference between the clean speech and noisy signal, and is defined as:

$$\text{PSM}(t, f) = \frac{|C(t, f)|}{|N(t, f)|} \cos(\phi_s(t, f) - \phi_x(t, f)) \quad (2.11)$$

where $\phi_c(t, f)$ and $\phi_n(t, f)$ are the phases of clean and noisy signals, respectively. Including phase information results in better enhancement performance, as it provides more accurate reconstruction of the clean signal.

5. Complex Ideal Ratio Mask (cIRM):

The cIRM is a technique introduced by Williamson et al.[38] in 2016 that aims to improve the separation of clean speech from noisy audio signals by leveraging both the magnitude and phase information of the signals. The main equations involved in the cIRM process can be summarized as follows:

In the Cartesian complex domain, the observed noisy signal $Y_{t,f}$ can be expressed in terms of its real and imaginary components. The equation describes the relationship between the clean speech $S_{t,f}$ and the noisy signal $Y_{t,f}$ as follows:

$$M = \frac{Y_r S_r + Y_i S_i}{Y_r^2 + Y_i^2} + i \frac{Y_i S_r - Y_r S_i}{Y_r^2 + Y_i^2} \quad (2.12)$$

Here, Y_r and Y_i represent the real and imaginary parts of the noisy signal, while S_r and S_i represent the real and imaginary parts of the clean speech. This formulation allows the model to capture the interaction between the components of both signals.

In evaluating spectrum masking methods, Wang et al.[39] discovered that spectral mask estimation significantly outperforms direct mapping in terms of improving voice quality and intelligibility. The IBM is one of the foundational targets for masking-based training and has demonstrated significant effectiveness in improving speech intelligibility for both individuals with normal hearing and those with hearing impairments[40]. To address the limitations of the IBM and its sharp binary labeling, the IRM[37] was introduced. The IRM is calculated by comparing the power spectrogram of clean speech to that of the noisy input. Another method for estimating signal magnitude is the SMM[41]. This technique emphasizes the signal's strength across various frequencies, aiding in the separation of the desired signal from background noise. Before the introduction of the PSM[42], existing masking techniques focused only on estimating the magnitude of signals. The PSM improves upon this by considering the relative phase difference between the clean and noisy signals. Experiments show that using phase information with PSM significantly increases the SDR compared to methods that enhance only the magnitude spectrum [43].

Williamson et al. [44] identified that analyzing the real and imaginary parts of

the clean speech spectrogram, rather than just the phase spectrogram, provides useful insights for supervised learning tasks. This insight led to the development of the cIRM, which recovers clean speech from noisy mixtures in the complex domain. However, cIRMs require complex-valued processing, making them computationally intensive and challenging for some network architectures. To address this issue, this study employs the IRM, which utilizes supervised learning while operating in the real domain, thus avoiding the complexities of larger network structures.

2.3 Research Gap

Recent advances in spectral masking techniques, as covered in the literature review, show clear improvements in speech quality and intelligibility compared to direct mapping methods. However, these techniques still face difficulties in performing well and adapting to different noisy environments, especially when noise levels are high (low SNR conditions). Leading methods like the Convolutional Recurrent Network (CRN), the Gated Convolutional Recurrent Network (GCRN), and Xiang’s masking-based approach have been successful but face challenges in managing complex time-frequency relationships and consistently achieving high intelligibility across different SNR levels. This highlights a research gap that needs to be addressed.

Additionally, recent developments in speech enhancement, particularly those using complex Ideal Ratio Mask (cIRM), involve computationally intensive processing and larger network architectures. This further emphasizes the need for alternative approaches that can maintain performance while reducing complexity.

This study aims to tackle these limitations by using the Ideal Ratio Mask (IRM) within two deep learning frameworks: a U-Net motivated architecture and a time-frequency transformer-based network(TF-TransNet).

Chapter 3: SE-UNet

In this chapter, a masking-based speech enhancement technique utilizing the U-Net architecture is introduced. The design of the proposed model is detailed, along with the training process and the necessary data preparation steps. Additionally, an analysis of the model’s output shows how well it works to improve speech intelligibility across various noise conditions compared to traditional methods.

3.1 Motivation

Initially created for medical image processing, the U-Net architecture [45], has gained popularity in SE due to its strong ability to reduce noise, making it suitable for applications like hearing aids and automatic speech recognition. Its effectiveness lies in its U-shaped structure, which consists of three main parts: an encoder, a decoder, and skip connections. The encoder compresses the input speech by down-sampling, extracting important features, while the decoder gradually restores the spatial dimensions through up-sampling, reconstructing the clean speech signal. A significant feature of the U-Net architecture is the skip connections, which directly transfer critical information from the encoder to the decoder, preserving details that are essential for accurate speech enhancement. This direct flow of information helps avoid the loss of important signal characteristics during reconstruction. In addition, the bottleneck between the encoder and decoder enables the model to capture high-level abstract features, contributing to more effective noise suppression. This combination of the U-Net structure provides a reliable way to improve speech quality and intelligibility in noisy environments.

3.2 Proposed Methodology

The Fig. 3.1 demonstrates the methodology of the proposed speech enhancement approach leveraging T-F masking technique. The process begins by taking a noisy input signal $X(t)$, which is converted from the time domain to the time-frequency domain using the STFT, resulting in its frequency representation $X(f)$. The

STFT also separates the signal into its magnitude $|X_{STFT}|$ and phase $\angle\theta_{X_{STFT}}$ components. Only the magnitude is used for further processing, while the phase is retained for later reconstruction.

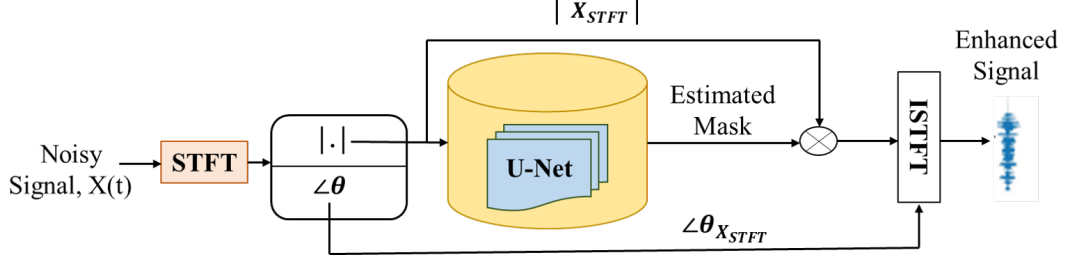


Figure 3.1: Framework of the proposed T-F masking technique-based speech enhancement method

The proposed attention based U-Net model is employed to estimate a spectral mask, designed to suppress the noise in the signal. Specifically, the IRM is used, which is computed as:

$$IRM(t, f) = \frac{Y^2(t, f)}{\sqrt{Y^2(t, f) + X^2(t, f)}}$$

where $Y^2(t, f)$ and $X^2(t, f)$ represent the energy of the clean signal and noise, respectively, for each time-frequency unit. The U-Net predicts this mask from the noisy magnitude $|X_{STFT}|$, and the estimated mask is applied to the noisy signal, yielding the enhanced magnitude spectrum

$$|Y(f)| = \hat{\text{Mask}}(f) \cdot |X_{STFT}|$$

The enhanced magnitude spectrum is then combined with the noisy phase information $\angle\theta_{X_{STFT}}$ to reconstruct the enhanced frequency-domain signal $\hat{Y}(f)$.

Finally, the inverse STFT (ISTFT) is applied to transform the enhanced frequency-domain signal back into the time domain, producing the enhanced time-domain output signal $\hat{Y}(t)$. This process effectively reduces noise while preserving speech quality, leveraging the time-frequency masking approach driven by the U-Net model.

3.2.1 Encoder-Decoder Architecture

This U-Net model in Fig. 3.2 for speech enhancement uses a structured encoder-decoder architecture with specific filter sizes to capture and reconstruct spectral features from noisy input. The input to the network is $X_0 \in \mathbb{R}^{H \times W \times C}$, where $H = 32$ represents the time steps, capturing the temporal resolution of the data,

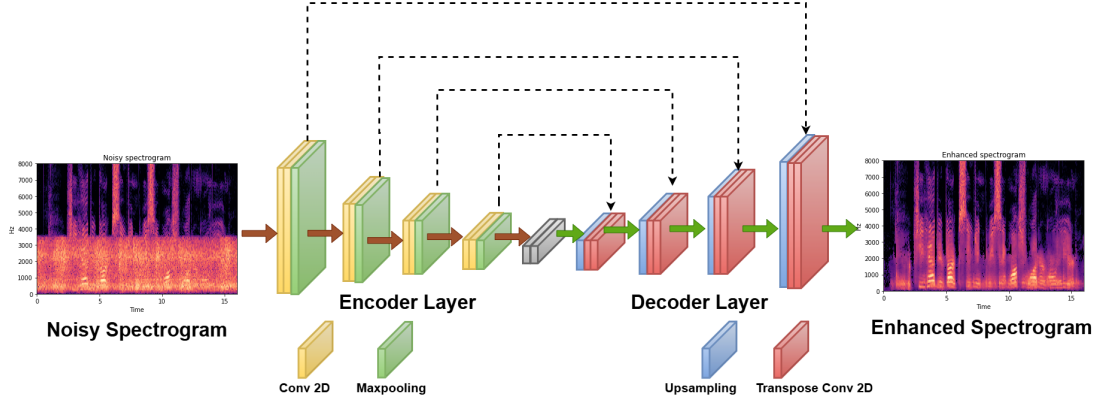


Figure 3.2: Architecture of the employed U-Net model for learning parameter optimization in the speech enhancement task, with distinct colors assigned to denote different operations in the encoder and decoder.

while $W = 129$ denotes the frequency bins, which reflect the spectral resolution of the input spectrogram and $C = 1$ indicates a single-channel input, typically a mono spectrogram. The encoder progressively reduces spatial dimensions while increasing filter depth from 64 to 512 (128, 256, and 512) filters with 3x3 kernels across four blocks, each followed by 2x2 MaxPooling for downsampling to capture essential features. , with a Dropout (0.5) layer in the final encoder block to reduce overfitting. Each encoder stage performs two convolutions, represented mathematically as:

$$X_{l,1} = \sigma(W_{l,1} * X_{l-1} + b_{l,1}), \quad (3.1)$$

$$X_{l,2} = \sigma(W_{l,2} * X_{l,1} + b_{l,2}), \quad (3.2)$$

where $W_{l,k}$ and $b_{l,k}$ are the learnable weights and biases, and σ is the ReLU activation function. These are followed by a max pooling operation:

$$P_l = \text{MaxPooling2D}(X_{l,2}, \text{pool_size} = (2, 2)), \quad (3.3)$$

which reduces the spatial dimensions by half while retaining the most significant features.

At the bottleneck, two additional convolutional layers with 1024 filters and 3x3 kernels capture abstract representations of the speech features are applied:

$$B_1 = \sigma(W_{b,1} * P_4 + b_{b,1}), \quad (3.4)$$

$$B_2 = \sigma(W_{b,2} * B_1 + b_{b,2}), \quad (3.5)$$

followed by dropout regularization:

$$B_{\text{drop}} = \text{Dropout}(B_2, \text{rate} = 0.5), \quad (3.6)$$

to reduce overfitting by randomly setting half of the features to zero.

The decoder, or expanding path, reconstructs the spatial resolution of the input through upsampling and skip connections. Upsampling is denoted as:

$$U_l = \text{UpSampling2D}(Z_{l+1}), \quad (3.7)$$

and skip connections concatenate these upsampled features with those from the corresponding encoder stage to restore spatial detail:

$$M_l = \text{Concatenate}([U_l, X_{\text{encoder}, l}]), \quad (3.8)$$

ensuring the preservation of spatial details from the encoder.

In each decoder stage, two convolutional operations refine the reconstructed features:

$$Z_{l,1} = \sigma(W_{l,1} * M_l + b_{l,1}), \quad (3.9)$$

$$Z_{l,2} = \sigma(W_{l,2} * Z_{l,1} + b_{l,2}), \quad (3.10)$$

Finally, the output layer generates a single-channel output tensor Y through a sigmoid activation function :

$$Y = \sigma_{\text{sigmoid}}(W_{\text{out}} * Z_{1,2} + b_{\text{out}}), \quad (3.11)$$

Table 3.1: Layer-wise output dimensions

Layer	Output Dimensions
Input	$32 \times 129 \times 1$
Encoder Block 1	$16 \times 64 \times 64$
Encoder Block 2	$8 \times 32 \times 128$
Encoder Block 3	$4 \times 16 \times 256$
Encoder Block 4	$2 \times 8 \times 512$
Bottleneck	$1 \times 4 \times 1024$
Decoder Block 4	$2 \times 8 \times 512$
Decoder Block 3	$4 \times 16 \times 256$
Decoder Block 2	$8 \times 32 \times 128$
Decoder Block 1	$16 \times 64 \times 64$
Output	$32 \times 129 \times 1$

3.2.2 Model Training and Optimization

Table 3.2 provides a summary of the optimized hyperparameters used for model training. The model employs the Adam optimizer with a learning rate of 0.001 and is compiled with mean squared error (MSE) as the loss function and evaluation metric. To enhance training efficiency, a model checkpoint is implemented, saving the best model based on validation loss. In addition, early stopping is used to avoid overfitting; training is stopped if the validation loss does not improve for 20 consecutive epochs. The model is trained over 50 epochs with a validation split of 20%, ensuring effective learning and optimization.

Table 3.2: Optimum hyperparameters value used

Hyperparameters	Optimum Value
Optimizer	adam
Learning rate	1×10^{-3}
Batch size	64
Epochs	5

3.2.2.1 Loss Function

Speech enhancement is one of the regression tasks that frequently employ the MSE loss function to measure the difference between expected and actual values. In the context of this task, the MSE measures how closely the enhanced speech signal generated by the model aligns with the target clean speech signal. The MSE is calculated using the following equation:

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (1)$$

where y_i represents the actual target value, $\hat{y}_i$ is the predicted value from the model, and n is the total number of samples. By minimizing the MSE, the model effectively reduces the error in predicting clean speech, thereby enhancing speech intelligibility and quality in noisy environments.

3.3 Experimental Setup

In this section, we outline the experimental setting in this section, with particular attention to the dataset and preprocessing procedures for the suggested speech enhancement models.

3.3.1 Corpora Description

This study evaluates the performance of the proposed SE model using the TIMIT dataset [46], which includes 6,300 phonetically transcribed sentences spoken by American English speakers. From this, a subset of 950 sentences spoken by 50 individuals was used for training. These sentences were artificially corrupted with eight types of environmental noises from the AURORA dataset (e.g., airport, babble, car, exhibition, restaurant, train station, street and speech shaped noise), applied at six different SNR levels: -10 dB, -5 dB, 0 dB, 5 dB and 10 dB. The aim was to simulate real-world conditions where background noise affects speech quality.

In the course of testing, the model is evaluated on both seen and unseen noise scenarios across the specified SNRs. Table 3.3 outlines the conditions for both seen and unseen noise types. The noise types and SNR values that are used to train the model are referred to as the "seen" condition. To assess the robustness of the proposed approach, the model is further evaluated under unseen conditions, specifically with two unfamiliar noises—airport and train noise—at SNRs of -7.5 dB and 7.5 dB, respectively.

Table 3.3: Evaluation Under Seen and Unseen Conditions for Different Noise Types and SNR Values in the Suggested Approach

Condition	Noise	SNR (dB)
Seen	SSN, Babble	-10, -5, 0, 5
Unseen	Airport, Train	-7.5, 7.4

All the speech and noise data were resampled at a standard rate of 16 kHz to ensure consistency across the dataset.

3.3.2 Tools and Libraries Used

Experiments are conducted on a GPU-facilitated platform using Google Colab-oratory. The NumPy (1.18.5) and pandas (1.1.4) libraries are used for data preparation and processing.. Audio feature extraction and manipulation are facilitated through the librosa library, while advanced signal processing tasks are managed using scipy.signal. The models are implemented using Keras (2.4.0) and TensorFlow (2.3.0), with various architectures such as CNNs and LSTMs utilized for the speech enhancement task.

The dataset is divided into training, validation, and test sets, where the training set is used for model learning, and hyperparameter tuning is conducted based on the validation set. The final evaluation of the trained models is performed

using the test set, allowing for an objective assessment of their performance in enhancing speech intelligibility and quality across various noise conditions.

3.4 Result Analysis

Using three objective metrics—STOI, PESQ, and SSNR—the results provides an overview of the efficacy of the suggested speech enhancement technique across various SNRs for both seen and unseen noise environments. Each metric compares the values of unprocessed noisy and enhanced speech, measuring several facets of speech quality and intelligibility.

Table 3.4: Performance of the Proposed Method in Terms of Speech Intelligibility and Quality

SNR	STOI				PESQ				SSNR			
	Seen Noise		Unseen Noise		Seen Noise		Unseen Noise		Seen Noise		Unseen Noise	
	Noisy	Enh.	Noisy	Enh.	Noisy	Enh.	Noisy	Enh.	Noisy	Enh.	Noisy	Enh.
-7.5	0.51	0.64	0.50	0.55	1.08	1.38	0.99	1.21	-7.92	-2.29	-7.74	-4.29
-5	0.57	0.69	0.56	0.61	1.19	1.57	1.10	1.38	-6.89	-1.03	-6.68	-3.26
0	0.69	0.80	0.68	0.73	1.41	1.89	1.31	1.72	-5.15	-1.05	-5.00	0.739
5	0.79	0.87	0.85	0.88	1.89	2.31	1.74	2.12	-2.05	1.56	-1.00	2.75
7.5	0.88	0.94	0.88	0.89	2.18	2.42	2.04	2.24	0.829	2.94	1.64	3.69
Mean	0.68	0.78	0.69	0.73	1.56	1.94	1.37	1.77	-4.12	0.90	-4.03	-1.32

3.4.1 Objective Metric Evaluation

From Table 3.4, it is evident that the STOI metric, which assesses intelligibility, shows a significant improvement in scores from the noisy speech to the enhanced speech across all SNR levels for both seen and unseen noise conditions. The mean STOI scores increase from 0.68 (noisy) to 0.78 (enhanced) in seen noise conditions and from 0.69 to 0.73 in unseen noise, indicating an overall enhancement in intelligibility due to the proposed method, with slightly better results for seen noise.

PESQ, a metric for perceived speech quality, also shows substantial improvement across all SNRs, with the enhanced speech consistently outperforming the noisy speech. The mean PESQ score improves from 1.56 to 1.94 for seen noise and from 1.37 to 1.77 for unseen noise. This indicates a noticeable enhancement in the perceived quality of the speech after applying the enhancement model, with a similar trend of higher improvement for seen noise conditions.

SSNR, which evaluates noise reduction by measuring the difference in SNR between noisy and enhanced speech, shows significant gains at higher SNRs (e.g., 5 dB and 7.5 dB). The mean SSNR improves from -4.12 to 0.90 for seen noise and from -4.03 to -1.32 for unseen noise. However, at lower SNRs (e.g., -7.5 dB),

SSNR remains negative, indicating that while noise reduction is achieved, it may not be sufficient in extremely noisy environments.

Overall, the proposed enhancement approach improves intelligibility and quality across various noise conditions, particularly in environments with moderate to high SNRs and for seen noise types. However, its performance in very low SNRs and unseen noise scenarios may be further refined for optimal effectiveness.

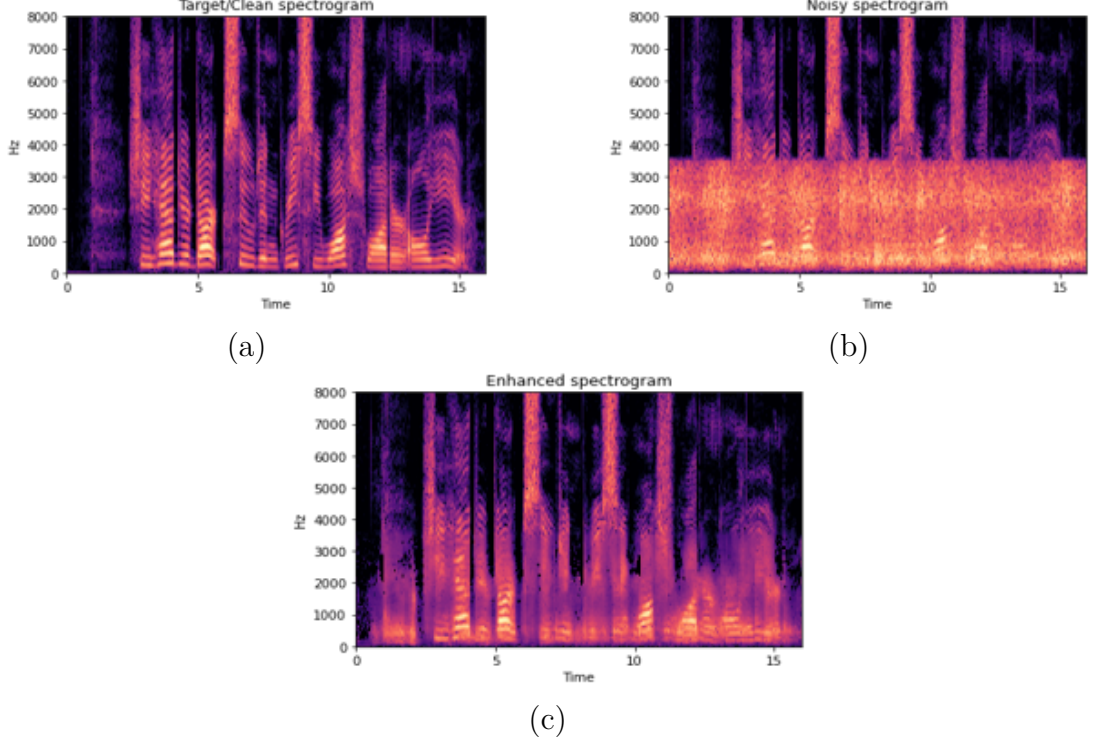


Figure 3.3: T-F representation of the clean signal utterance is represented (a), followed by the spectrogram of the noisy one (b) and the enhanced spectrogram (c).

3.4.2 Spectrogram Analysis

Figure 3.4.1 presents the spectrograms for clean, noisy, and enhanced signals. It is clear from the figure that the spectral masking technique significantly enhances the suppression of non-stationary noises in the noisy spectrogram. However, the enhanced spectrogram exhibits several artifacts that impact its quality. Residual noise persists, especially in silent or low-energy regions, as seen in Fig. 3.4.1, which compromises intelligibility. Additionally, harmonic smearing occurs, where the distinct vertical stripes representing harmonic structures in the clean spectrogram become less defined, negatively affecting clarity. Some subtle spectral details are also lost, reducing the naturalness of the enhanced speech.

3.4.3 Comparison with Mapping Strategies

Table 3.5: Comparison of Speech Enhancement Models Based on STOI, PESQ, SSNR Metrics

Models	STOI	PESQ	SSNR
Unprocessed	0.68	1.46	-4.08
CNN [47]	0.71	1.58	-0.22
Proposed SE-UNet (Mapping Strategies)	0.70	1.51	-2.38
Proposed SE-UNet (Masking Strategies)	0.76	1.86	-0.21

Table 3.5 presents a comparison of the proposed speech enhancement model against a noisy mixture and an existing mapping based CNN architecture[47]. The noisy mixture achieves a STOI score of 0.68, while the existing CNN[47] improves this to 0.71. The proposed mapping-based approach scores 0.70, and the masking-based method outperforms all with a score of 0.76, indicating better intelligibility. For PESQ, the noisy mixture scores 1.46, which the CNN [47] improves to 1.58. The proposed models score 1.51 (mapping-based) and 1.86 (masking-based), reflecting enhanced speech quality. In SSNR, the noisy mixture has a score of -4.08, while the CNN improves it to -0.22. The proposed mapping-based model scores -2.38, and the masking-based model again excels with a score of -0.21. The masking-based model outperforms both the noisy mixture and the existing CNN [47] across all evaluated metrics, highlighting its effectiveness in improving speech quality and comprehensibility in challenging noisy environments.

3.5 Summary

This chapter assesses the proposed speech enhancement method’s ability to enhance voice quality and intelligibility under various noise levels. The findings indicate that the model performs particularly well in environments with moderate to high SNRs and for seen noise types. However, the results also suggest that the model’s effectiveness diminishes in scenarios with very low SNRs and unseen noise types. These limitations highlight the potential for further refinement of the approach to optimize its performance in challenging acoustic environments. Detailed analysis and recommendations for improvement will be discussed in Chapter 4.

Chapter 4: TF-TransNet

This chapter presents a new approach for enhancing speech signals, addressing the challenges of previous SE-UNet posed by very low SNRs and unseen noise types. To overcome the limitations, we employ a TF-TransNet in this study. The model incorporates a TFA mechanism alongside a Time-Frequency Transformer, organized within an encoder-decoder architecture to effectively reconstruct and improve speech signals degraded by noise in the spectral domain.

4.1 Motivation

By leveraging a TFA mechanism, TF-TransNet aims to focus on the most relevant components of the audio signal, thereby enhancing critical speech features while minimizing the impact of noise. This study focuses on the encoder component of the transformer architecture, particularly in the bottleneck layer. Research indicates that dual-path attention-based architectures [48, 49] excel in speech enhancement and separation tasks [50] due to their ability to model both temporal (time-based) and spectral (frequency-based) dependencies in speech signals. The TF Transformer layer used in [51] is designed to effectively capture long-range dependencies in speech signals over both time and frequency domains. Building on the approach in [51], this study implements a TF Transformer as a bottleneck layer, as illustrated in Figure 4.3. The architecture includes a Time-Transformer layer, which focuses on temporal dependencies, and a Frequency-Transformer layer, which captures frequency interactions within the data. The encoder-decoder architecture allows for a comprehensive analysis of both the temporal and spectral dimensions of the speech signal, facilitating the reconstruction of a clearer, more intelligible output.

4.2 Proposed Methodology

The architecture of the proposed TF-TransNet is shown in Fig. 4.1, which is designed to enhance speech signals from noisy audio using spectral masking as a core technique. The process begins by applying the STFT to the noisy signal $X(t)$ to extract the short-time complex spectrum $X_{t,f} \in \mathbb{C}^{T \times F}$, where T denotes

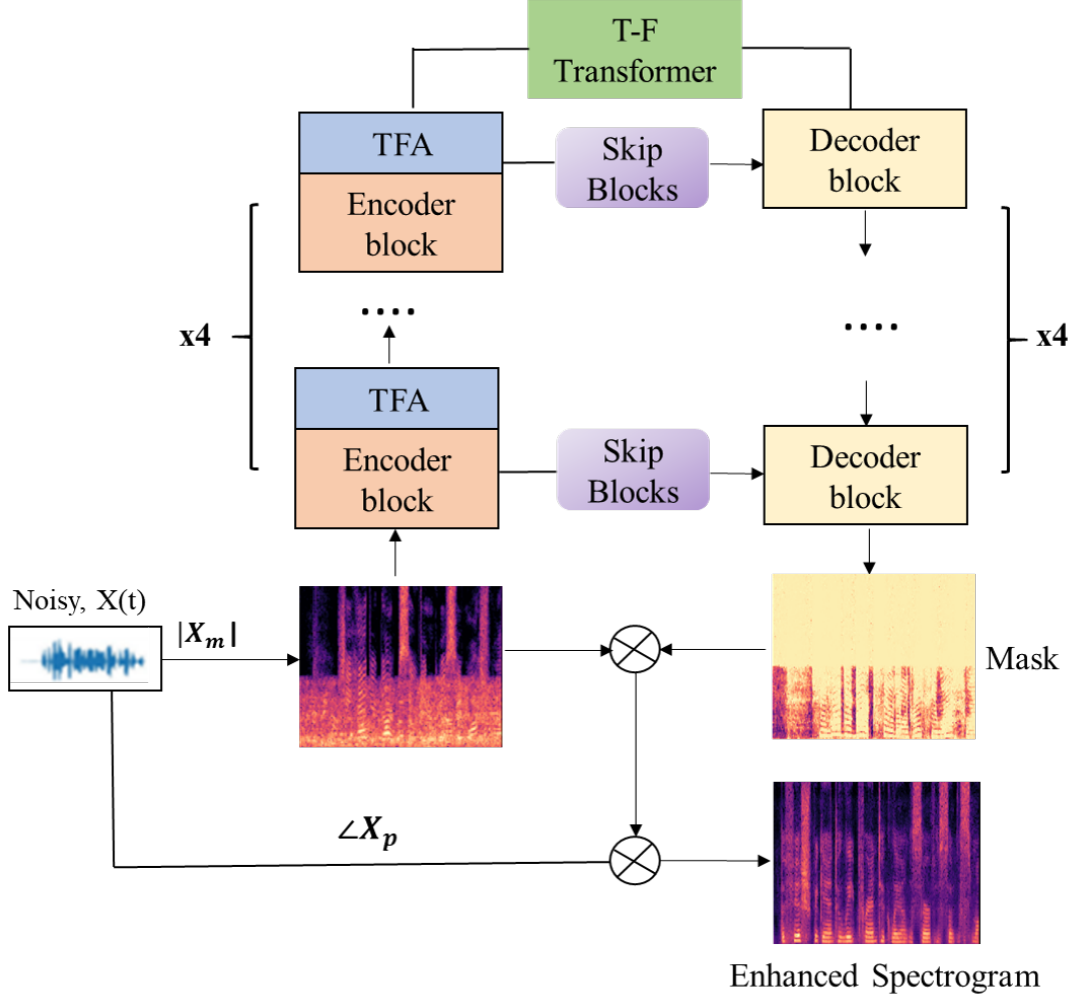


Figure 4.1: Block diagram illustrating the proposed TF-TransNet SE method, with spectral masking as its ceore technique.

the number of time frames and F represents the number of frequency bins. The magnitude spectrum $|X_m| \in \mathbb{R}^{T \times F}$ and the wrapped phase spectrum $\angle X_p \in \mathbb{R}^{T \times F}$ are subsequently calculated, such that $X = X_m \cdot e^{jX_p}$.

The architecture consists of multiple encoder blocks with TFA that focus on significant time-frequency features, effectively enhancing relevant portions of the signal while suppressing noise. The processed spectrogram information from the encoder is directed to a bottleneck layer, where TF-transformer is utilized to achieve improved noise suppression. Skip blocks are integrated to facilitate information flow between layers, allowing for the combination of features from earlier and deeper layers, thereby retaining essential details.

The decoder blocks reconstruct the enhanced output from the encoded features, generating a mask applied to the magnitude spectrogram of the noisy input.

Let $|S(t, f)|$ represent the clean signal magnitude at time t and frequency f ,

while $|X(t, f)|$ denotes the noisy magnitude at the same time and frequency. The IRM, denoted as $M \in \mathbb{R}^{T \times F}$, at a specific time t and frequency f is computed as follows:

$$\text{IRM}(t, f) = \frac{S^2(t, f)}{S^2(t, f) + X^2(t, f)} \quad (4.1)$$

The enhanced magnitude spectrum is derived through element-wise multiplication of the noisy signal magnitude spectrum and the predicted mask $\hat{\text{IRM}}$ (where $\hat{M} \in \mathbb{R}^{T \times F}$):

$$|\hat{S}(t, f)| = \hat{\text{IRM}}(t, f) \cdot |X(t, f)| \quad (4.2)$$

Finally, the enhanced time-domain signal $\hat{S}(t)$ is reconstructed by computing the iSTFT using the enhanced magnitude spectrum $|\hat{S}(t, f)|$ and the phase of the noisy signal $\angle X(t, f)$:

$$\hat{S}(t) = \text{iSTFT} \left\{ |\hat{S}(t, f)| e^{j\angle X(t, f)} \right\} \quad (4.3)$$

4.2.1 Encoder Decoder Architecture

The encoder comprises a series of convolutional blocks, each equipped with two convolutional layers that use ReLU activation functions and dilation rates. By capturing multi-scale temporal and spectral patterns, these dilation rates enable the network to expand the receptive field without adding more parameters. After each convolutional layer, a max-pooling operation is performed, reducing the spatial dimensions of the feature maps while retaining key high-level features. This enables the model to downsample the input while progressively focusing on relevant audio features.

Block 1 has two convolutional layers with 32 filters each, using a (3, 3) kernel size, ReLU activation, and a dilation rate of (1, 2), followed by a TFA mechanism and max-pooling. Block 2 includes two layers with 64 filters and similar dilation, downsampled through max-pooling. Block 3 has two layers with 128 filters and max-pooling, while Block 4 uses 256 filters with similar configuration, and the output is reshaped for the bottleneck layer.

The decoder architecture is designed to reconstruct the input resolution through four sequential upsampling blocks, which reverse the operations of the encoder. Each block consists of two transposed convolutional layers, which are responsible for refining features and increasing spatial resolution. The layers utilize decreasing numbers of filters as follows: 256 filters in the first block, 128 filters in the second, 64 filters in the third, and 32 filters in the fourth. Upsampling

is performed using 2D upsampling layers, and skip connections are implemented by cropping and concatenating corresponding feature maps from the encoder to retain high-resolution information. This strategy ensures that critical features from the encoding phase are preserved during decoding. The final output is produced by a transposed convolutional layer with a single filter, generates the output mask, which is applied to the noisy spectrogram to produce the enhanced signal.

4.2.2 Time - Frequency Attention (TFA)

The TFA mechanism embedded in the encoder is designed to process feature maps by taking into account both their temporal and frequency characteristics. The input to the TFA mechanism is a tensor $X \in \mathbb{R}^{T \times F \times C}$, where T represents the number of time steps, F represents the number of frequency bins, and C is the number of channels. This input tensor is then reshaped into two distinct branches: a frequency branch (FB) and a temporal branch (TB), each processing different aspects of the input data to extract relevant features, as illustrated in Fig. 4.2. The input is transformed channel-wise in both branches using 1×1 convolution layers followed by ReLU activation. These transformations yield query (T_q, F_q), key (T_k, F_k), and value (T_v, F_v) representations for the temporal (T) and frequency (F) branches, respectively.

Frequency Branch (FB): The frequency branch operates on the reshaped input $X_f \in \mathbb{R}^{T \times F \times C}$. Query (F_q), Key (F_k), and Value (F_v) are generated through 1×1 convolutions:

$$F_q = \text{Conv2D}_{1 \times 1}(X_f), \quad F_k = \text{Conv2D}_{1 \times 1}(X_f), \quad F_v = \text{Conv2D}_{1 \times 1}(X_f) \quad (4.4)$$

The attention map A_f is computed via a softmax function applied to the matrix multiplication of the query and key:

$$A_f = \text{softmax}(F_q \cdot F_k^T) \quad (4.5)$$

The attention output is then computed as a weighted sum of the values F_v features, thereby highlighting important features based on their relevance at each time-frequency position:

$$F_{\text{out}} = A_f \cdot F_v \quad (4.6)$$

The resulting weighted features, F_{out} , are reshaped to match the original feature map dimensions for subsequent stages.

Temporal Branch (TB):

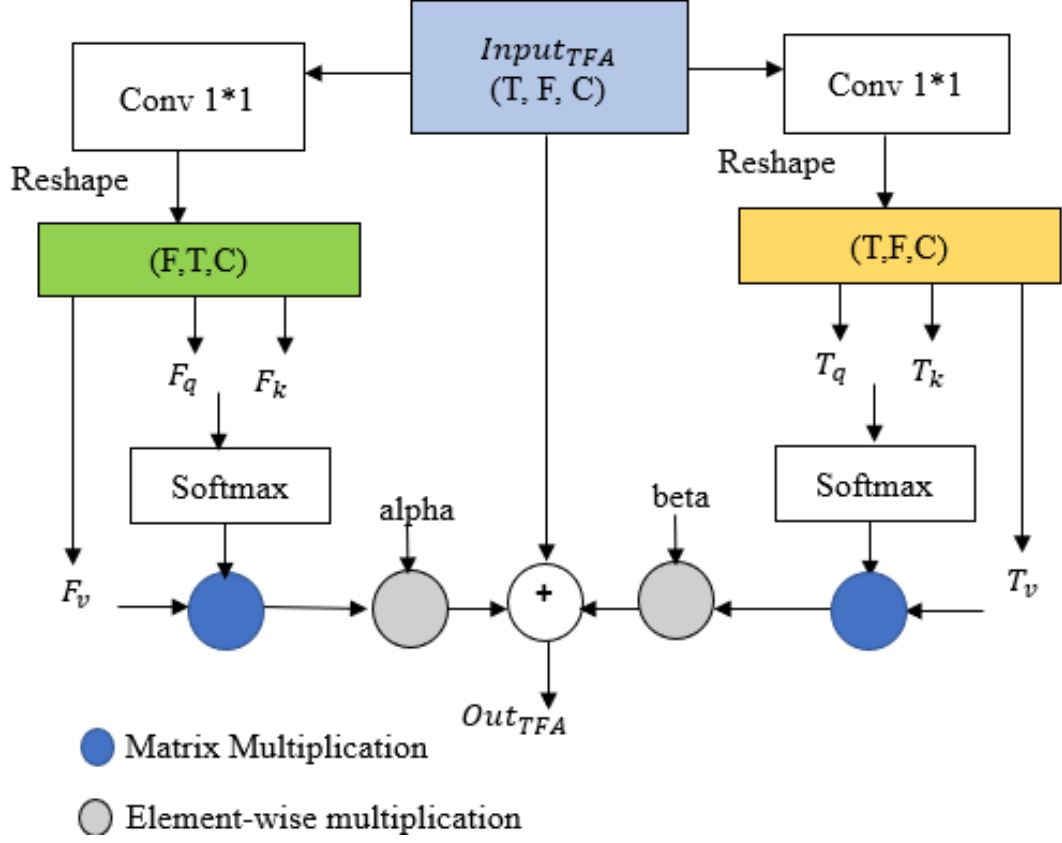


Figure 4.2: Time-Frequency Attention mechanism employed in the Encoder. The input to the attention mechanism, $Input_{TFA}$ has dimensions (T, F, C) , where T represents number of frames, F represents frequency bins, and C represents the number of channels.

The temporal branch works on the reshaped input $X_t \in \mathbb{R}^{F \times T \times C}$, where the dimensions of time and frequency are swapped. Similar to the frequency branch, the following operations are performed:

TB similarly transforms the input through three convolutional layers along the temporal dimension to generate query (T_q), key (T_k), and value (T_v) representations.

$$T_q = \text{Conv2D}_{1 \times 1}(X_t), \quad T_k = \text{Conv2D}_{1 \times 1}(X_t), \quad T_v = \text{Conv2D}_{1 \times 1}(X_t) \quad (4.7)$$

These representations are further transformed using fully connected (Dense) layers, producing query and key vectors suitable for calculating attention. The Dense layers, employing ReLU activation, project the reshaped representations (T_q, T_k) and T_v to a lower-dimensional space, facilitating efficient attention computation. As with the FB, the attention map A_t is computed by taking the dot product between the query and key vectors, followed by a softmax function:

$$A_t = \text{softmax}(T_q \cdot T_k^T) \quad (4.8)$$

The resulting attention weights are applied to the value vectors, producing a temporal-focused output, T_{out} :

$$T_{out} = A_t \cdot T_v \quad (4.9)$$

Finally, the output is reshaped back to the original input shape:

$$t_{out_reshaped} \in \mathbb{R}^{T \times F \times C} \quad (4.10)$$

The outputs from both the frequency branch F_{out} and the temporal branch T_{out} are then combined with the original input feature map using the parameters α and β . These weights determine the contribution of each branch to the final output and are initially set to 0.5, but can be adjusted during training to optimize performance.

$$\text{Out}_{\text{TFA}} = \text{inputs} + \alpha \cdot F_{out} + \beta \cdot T_{out} \quad (4.11)$$

4.2.3 T-F Transformer Layer

The output from the encoder's downsampling layers is represented as feature maps, mathematically denoted by $X_{enc} \in \mathbb{R}^{B \times T \times F \times C}$. The Time-Transformer and Frequency-Transformer modules enhance the input features by leveraging temporal and frequency relationships using a combination of MHSA [33] to capture relationships across different parts of the speech, along with a LSTM-based recurrent network to learn local information within each time step or frequency component . T

Time-Transformer

For the Time-Transformer layer, this input tensor is reshaped into $X' \in \mathbb{R}^{(B \times F) \times T \times C}$ to focus on temporal relationships between different time steps. The reshaped tensor is processed through MHSA, which computes the attention output as:

$$A_{time} = \text{Dropout}(\text{MHSA}(X', X')). \quad (4.12)$$

A residual connection is applied, followed by layer normalization:

$$Y_{time} = \text{LayerNorm}(X' + A_{time}). \quad (4.13)$$

Subsequently, an LSTM layer is applied to capture local temporal patterns,

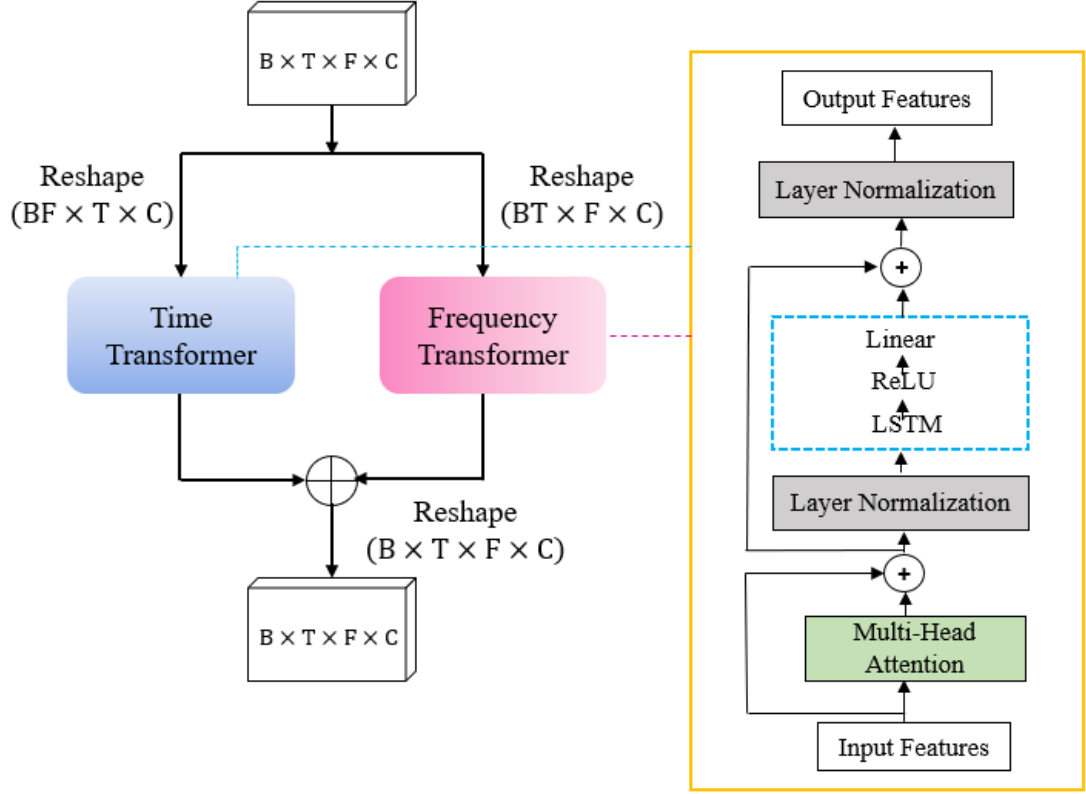


Figure 4.3: Architecture of the TF-Transformer layer showcased in Fig. 4.1, consisting of a Time-Transformer and a Frequency-Transformer. Both maintain identical architectures.

yielding:

$$L_{\text{time}} = \text{Dropout}(\text{LSTM}(Y_{\text{time}})). \quad (4.14)$$

Another residual connection and layer normalization result in:

$$Z_{\text{time}} = \text{LayerNorm}(Y_{\text{time}} + L_{\text{time}}). \quad (4.15)$$

Finally, the tensor is reshaped back to:

$$Z_{\text{time}} \in \mathbb{R}^{B \times T \times F \times C}. \quad (4.16)$$

Frequency-Transformer

In the Frequency-Transformer, the input tensor is reshaped to $X' \in \mathbb{R}^{(B \times T) \times F \times C}$ to capture frequency relationships. MHSA is applied to compute:

$$A_{\text{freq}} = \text{Dropout}(\text{MHSA}(X', X')). \quad (4.17)$$

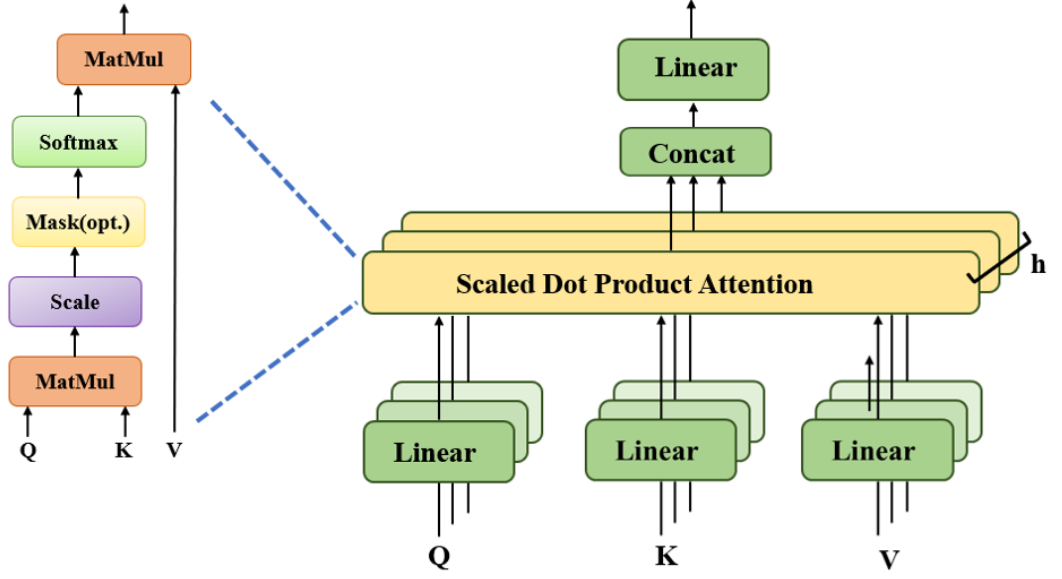


Figure 4.4: Multihead attention mechanism that is employed in the bottleneck layer of the proposed architecture.

A residual connection and layer normalization produce:

$$Y_{\text{freq}} = \text{LayerNorm}(X' + A_{\text{freq}}). \quad (4.18)$$

A LSTM layer is then employed to extract frequency dependencies:

$$L_{\text{freq}} = \text{Dropout}(\text{LSTM}(Y_{\text{freq}})). \quad (4.19)$$

Finally, another residual connection and normalization yield:

$$Z_{\text{freq}} = \text{LayerNorm}(Y_{\text{freq}} + L_{\text{freq}}), \quad (4.20)$$

and the tensor is reshaped back to:

$$Z_{\text{freq}} \in \mathbb{R}^{B \times T \times F \times C}. \quad (4.21)$$

These modules effectively refine the feature representations, with the Time-Transformer focusing on temporal dependencies and the Frequency-Transformer emphasizing frequency relationships.

4.2.3.1 Multi Head Self Attention

Scaled dot-product attention is a self-attention mechanism that calculates the importance of each position in an input sequence relative to a specific "query"

position. This importance score is used to weight the corresponding "value" from the sequence. The final output is obtained by summing the weighted values. Multi-head attention, depicted in Figure 4.4, leverages multiple instances of scaled dot-product attention in parallel, allowing the model to capture diverse aspects of the relationships within the sequence. The MHSA component with M heads allows the model to simultaneously focus on global information from different representation subspaces at various positions. The MHSA is formulated as:

$$\text{MHSA}(Q, K, V) = \text{Concat}(\text{head}_1, \dots, \text{head}_M)W_O \quad (4.22)$$

where each head is defined as:

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \quad (4.23)$$

and the attention mechanism is computed as:

$$\text{Attention Scores}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (4.24)$$

Here, Q , K , and V are the query, key, and value matrices, W_i^Q , W_i^K , W_i^V , and W_O are learned weight matrices, and d_k is the dimensionality of the key vectors. For the proposed network, number of heads, $M=2$.

4.3 Experimental Setup

In this section, we describe the experimental setup, focusing on the dataset and preprocessing steps for the proposed speech enhancement models.

4.3.1 Corpora Description

The dataset used for evaluation is the same TIMIT dataset [46] as outlined in Chapter 3. However, the conditions for the network have been modified to incorporate both seen and unseen noise conditions. A detailed summary of these conditions is presented in table 4.1. A total of 14,000 samples are generated by augmenting 500 original samples with 7 different noise types and 5 distinct SNR levels, with 80% used for training and 20% for validation. The unseen condition consists of 50 samples, each subjected to the noise types Restaurant and Train, with 5 distinct SNR levels (-7.5 dB, -2.5 dB, 2.5 dB, 7.5 dB, 15 dB), resulting in a total of 500 samples for testing.

Table 4.1: Seen and Unseen Conditions for Various Noise Types and SNR Values of the Proposed Work

Condition	Noise	SNR (dB)
Seen	Airport, Babble, Car, Exhibition, Station, Street, SSN	-10, -5, 0, 5, 10
Unseen	Restaurant, Train	-7.5, -2.5, 2.5, 7.5, 15

4.3.2 Tools and Libraries Used

Experiments are conducted on a GPU-facilitated platform using Google Colab. Data processing and preparation are performed with the pandas (1.1.4) and NumPy (1.18.5) libraries. Audio feature extraction and manipulation are facilitated through the librosa library, while advanced signal processing tasks are managed using scipy.signal. The models are implemented using Keras (2.4.0) and TensorFlow (2.3.0), with various architectures such as CNNs and LSTMs utilized for the speech enhancement task.

The dataset is divided into training, validation, and test sets, where the training set is used for model learning, and hyperparameter tuning is conducted based on the validation set. The final evaluation of the trained models is performed using the test set, allowing for an objective assessment of their performance in enhancing speech intelligibility and quality across various noise conditions.

4.4 Results

This section examines the performance of the proposed TF-TransNet architecture on enhancing speech. The assessment includes a comprehensive analysis of results obtained from various objective metrics, particularly emphasizing speech intelligibility and quality. We conduct a comparative evaluation of scores under different conditions, taking into account both known and novel noise types, as well as varying Signal-to-Noise Ratios (SNRs)

4.4.1 Evaluating Performance through Objective Metrics

The results summarized in the table 4.2 provides a detailed comparison of the enhanced signal from the proposed SE method and a noisy signal across various SNR levels and different types of noise, specifically babble, car, and speech-shaped noise.

Table 4.2: Performance Comparison of Noisy and Proposed Methods Across Various Noise Types and SNR Levels

SNR	Method	STOI			PESQ			SISDR		
		Babble	Car	SSN	Babble	Car	SSN	Babble	Car	SSN
-5 dB	Noisy	0.52	0.55	0.59	1.17	1.09	1.22	-5.06	-5.05	-4.87
	Proposed	0.71	0.72	0.74	1.56	1.56	1.64	2.37	2.55	2.62
0 dB	Noisy	0.65	0.66	0.69	1.43	1.37	1.56	0.04	-0.03	-0.03
	Proposed	0.80	0.81	0.81	1.97	1.98	2.00	6.84	6.85	6.10
5 dB	Noisy	0.79	0.79	0.79	1.80	1.73	1.92	4.98	5.08	5.01
	Proposed	0.88	0.88	0.87	2.35	2.39	2.37	10.42	11.05	10.47

The proposed method consistently outperforms the noisy signal, particularly in terms of speech intelligibility, STOI scores. For example, at -5 dB SNR, the STOI score increases from 0.52 for the noisy method to 0.71 for the proposed method, reflecting a significant improvement in speech clarity and understandability. Similar trends are observed at 0 dB and 5 dB SNR, indicating enhanced clarity and understandability of speech. In terms of speech quality, the PESQ scores also exhibit notable enhancements, rising from 1.17 (noisy) to 1.56 (proposed) at -5 dB SNR, with similarly improved scores at higher SNR levels.

Additionally, the SISDR scores indicate a substantial reduction in distortion, moving from -5.06 for the noisy method to 2.37 for the proposed method at -5 dB SNR, with consistent improvements at 0 dB and 5 dB. Overall, these results confirm that the proposed speech enhancement method significantly improves speech intelligibility, quality, and signal-to-distortion ratio across different noise types and SNR levels, underscoring its effectiveness in enhancing speech clarity in challenging auditory environments.

4.4.2 Ablation Study of the Proposed Framework

Table 4.3 shows the results of an ablation study that investigates the effects of different components in the proposed TF-TransNet speech enhancement model. Adding the TFA mechanism along with the encoder and decoder leads to slight improvements compared to unprocessed speech. When a temporal transformer (T-Trans.) is included, performance improves even more, while a frequency transformer (F-Trans.) provides similar results. Combining TFA with the temporal transformer and decoder results in notable enhancements, and a similar combination with the frequency transformer also works well. When both temporal and frequency transformers are used with the encoder and decoder, there are significant performance gains. The complete TF-TransNet model, which includes TFA, the encoder, both transformers, and the decoder, achieves the best performance (STOI: 0.86, PESQ: 2.40). The TFA mechanism helps focus on important time-frequency features, while the combination of both transformers captures

Table 4.3: Ablation Study of the Proposed Network in Terms of Objective Intelligibility and Quality

Method	STOI	PESQ
Unprocessed	0.67	1.44
TFA+Encoder+Decoder	0.69	1.76
Encoder+T-Trans.+Decoder	0.73	1.85
Encoder+F-Trans.+Decoder	0.71	1.88
TFA+Encoder+T-Trans.+Decoder	0.79	1.91
TFA+Encoder+F-Trans.+Decoder	0.76	1.91
Encoder+T-F Trans.+Decoder	0.82	2.13
TF-TransNet	0.86	2.4
(TFA+Encoder+TF Trans.+Decoder)		

connections across time and frequency, greatly improving speech intelligibility and quality.

4.4.3 Comparison with Existing Networks

As a comparative analysis of the proposed network’s performance , this study compares the proposed network in table 4.4 with three established SOTA approaches: CRN[28],GCRN [52] and a making-inpainting approach in Xiang et al. [53]. The CRN algorithm integrates CNN and RNN to leverage the strengths of both CNNs for local feature extraction and RNNs for sequential modeling, potentially leading to improved performance. GCRN combines CNNs and RNNs with gating mechanisms which allows the network to selectively process and retain relevant information over extended time periods. This selective learning capability makes GCRN particularly adept at handling sequential data characterized by long-term dependencies and recurring temporal patterns. The fusion of these two network types empowers GCRN to effectively process complex sequences and extract valuable insights from them. Xiang et al. [53] employs a two-stage method for the speech enhancement task. The first stage involves calculating a binary mask, which is subsequently applied to the noisy spectrogram. In the second stage, the enhanced spectrogram is estimated based on the masked spectrogram using CNN architecture.

For the Seen condition, TF-TransNet achieves the highest scores across all metrics. In STOI, it scores 0.92 in high SNR, 0.85 in medium, and 0.77 in low SNR. This results in an average score of 0.85 , compared to an average of 0.71 for the noisy signal. Similarly, for PESQ, TF-TransNet attains a high score of

Table 4.4: Performance Comparison with Existing Models in Terms of Objective Intelligibility, Quality, and Distortion Scores for Three SNR Groups

Condition	Method	STOI				PESQ				SISDR			
		SNR Range				SNR Range				SNR Range			
		High	Med.	Low	Avg.	High	Med.	Low	Avg.	High	Med.	Low	Avg.
Seen	Noisy	0.85	0.71	0.59	0.71	1.99	1.53	1.21	1.57	7.50	1.26	-3.74	1.67
	CRN	0.87	0.77	0.65	0.76	2.15	1.69	1.56	1.80	6.00	3.30	0.16	3.15
	GCRN	0.89	0.79	0.67	0.79	2.20	1.78	1.59	1.86	7.85	4.29	1.58	4.57
	Xiang et al.[53]	0.90	0.81	0.69	0.80	2.32	1.90	1.63	1.95	6.57	5.83	2.80	5.07
	TF-TransNet	0.92	0.85	0.77	0.85	2.87	2.58	2.00	2.48	13.55	8.36	5.26	9.06
Unseen	Noisy	0.84	0.73	0.61	0.73	1.54	1.43	1.12	1.36	7.49	1.19	-3.82	1.62
	CRN	0.86	0.74	0.61	0.74	1.56	1.48	1.25	1.43	5.76	2.20	-2.19	1.92
	GCRN	0.85	0.77	0.65	0.76	1.64	1.67	1.29	1.53	5.76	3.28	0.26	3.10
	Xiang et al.[53]	0.86	0.74	0.61	0.74	1.73	1.89	1.34	1.65	7.80	4.32	1.98	4.70
	TF-TransNet	0.90	0.81	0.72	0.81	2.63	2.35	1.78	2.25	13.16	7.79	3.50	8.15

2.87 in high SNR, 2.58 in medium, and 2.00 in low SNR, with an average of 2.48, surpassing the performance of all other methods. In SISDR, TF-TransNet shows remarkable improvement with scores of 13.55, 8.36, and 5.26 for high, medium, and low SNR, respectively, resulting in an average score of 9.06.

In the Unseen condition, TF-TransNet continues to outperform the other methods, with scores of 0.90 (high SNR), 0.81 (medium), and 0.72 (low) in STOI, yielding an average of 0.81. The PESQ scores are 2.63, 2.35, and 1.78, with an average of 2.25, indicating strong speech quality enhancement. For SISDR, TF-TransNet also led with 13.16 (high), 7.79 (medium), and 3.50 (low), achieving an average of 8.15.

Table 4.5: Comparison with other SOTA models to Assess Objective Intelligibility, Quality, and Distortion Scores

Method	STOI	PESQ	SISDR
Unprocessed	0.67	1.4	0.37
CRN	0.72	1.58	1.72
GCRN	0.78	1.83	3.59
Xiang et al.[53]	0.82	2.00	3.78
TF-TransNet	0.86	2.40	8.60

To validate the robustness of the proposed TF-TransNet model, the study presents average objective scores obtained under various noise conditions, as shown in table 4.5. The experiments include seven different signal-to-noise ratios (SNRs) (-7.5, -5, -2.5, 0, 2.5, 5, and 10 dB) and eight types of noise, consisting of five familiar and three unfamiliar noise types. The results demonstrate that the TF-TransNet model significantly improves performance compared to both unprocessed audio and other baseline models.

In terms of STOI, which measures speech intelligibility, TF-TransNet achieves a substantial 28.36% improvement over unprocessed audio. It also outperforms the baseline models, showing a 19.44% improvement over the CRN model, 10.26% over the GCRN model, and 4.88% over the method proposed by Xiang et al.[53].

This clearly demonstrates TF-TransNet’s superior ability to enhance speech intelligibility.

For PESQ, which evaluates the perceived quality of speech, TF-TransNet shows a remarkable 71.43% improvement over unprocessed audio. It also displays significant gains over other models, improving PESQ by 51.90% compared to CRN, 31.15% over GCRN, and 20.00% over Xiang et al.’s [53] method. This highlights the model’s effectiveness in improving speech quality.

In terms of SISDR, which measures distortion reduction, TF-TransNet achieves an impressive 8.23 dB improvement over unprocessed audio. It also surpasses the CRN model by 6.88 dB, the GCRN model by 5.01 dB, and Xiang et al.’s method by 4.82 dB. These results collectively validate TF-TransNet’s superior performance across different noise conditions, demonstrating its robustness in enhancing speech intelligibility, quality, and reducing distortion.

To support the findings, Fig.4.5 compares spectrograms of noisy, clean, and enhanced speech samples processed by three existing state-of-the-art models and the proposed TF-TransNet model. The CRN (c) and GCRN (d) models show some noise reduction, but they struggle to maintain the harmonic structures of the speech. The model by Xiang et al. [53] (e) performs better, offering improved noise reduction while preserving more of the speech’s harmonic features. However, the proposed TF-TransNet model (f) outperforms all others, effectively reducing noise while closely preserving the harmonic structures, making it nearly indistinguishable from the clean speech sample. The colored boxes in the figure clearly illustrate that TF-TransNet excels in enhancing speech quality by achieving a good balance between noise reduction and harmonic preservation.

Another analysis of the spectrograms reveals additional insights into the performance of the models. The Fig. 4.6 provides a detailed comparison of clean, noisy, and enhanced speech signals, along with the differences between these signals at an SNR of 2.5 dB. The clean spectrogram displays the undistorted, original speech signal, serving as the reference. The noisy spectrogram shows the speech signal corrupted by noise at 2.5 dB SNR, where the noise is clearly visible as irregularities in the spectrogram. The enhanced spectrogram, processed by TF-TransNet, shows a significant reduction in noise and a restoration of speech features, closer to the clean signal. The difference between clean and enhanced spectrogram highlights the improvements made by the enhancement process, particularly in regions where the harmonic structures have been successfully recovered. The difference between noisy and enhanced spectrogram further quantifies the reduction in noise, emphasizing that the TF-TransNet model excels in both noise suppression and harmonic preservation. The figure demonstrates the efficacy of the enhancement method in reducing noise and recovering speech clarity,

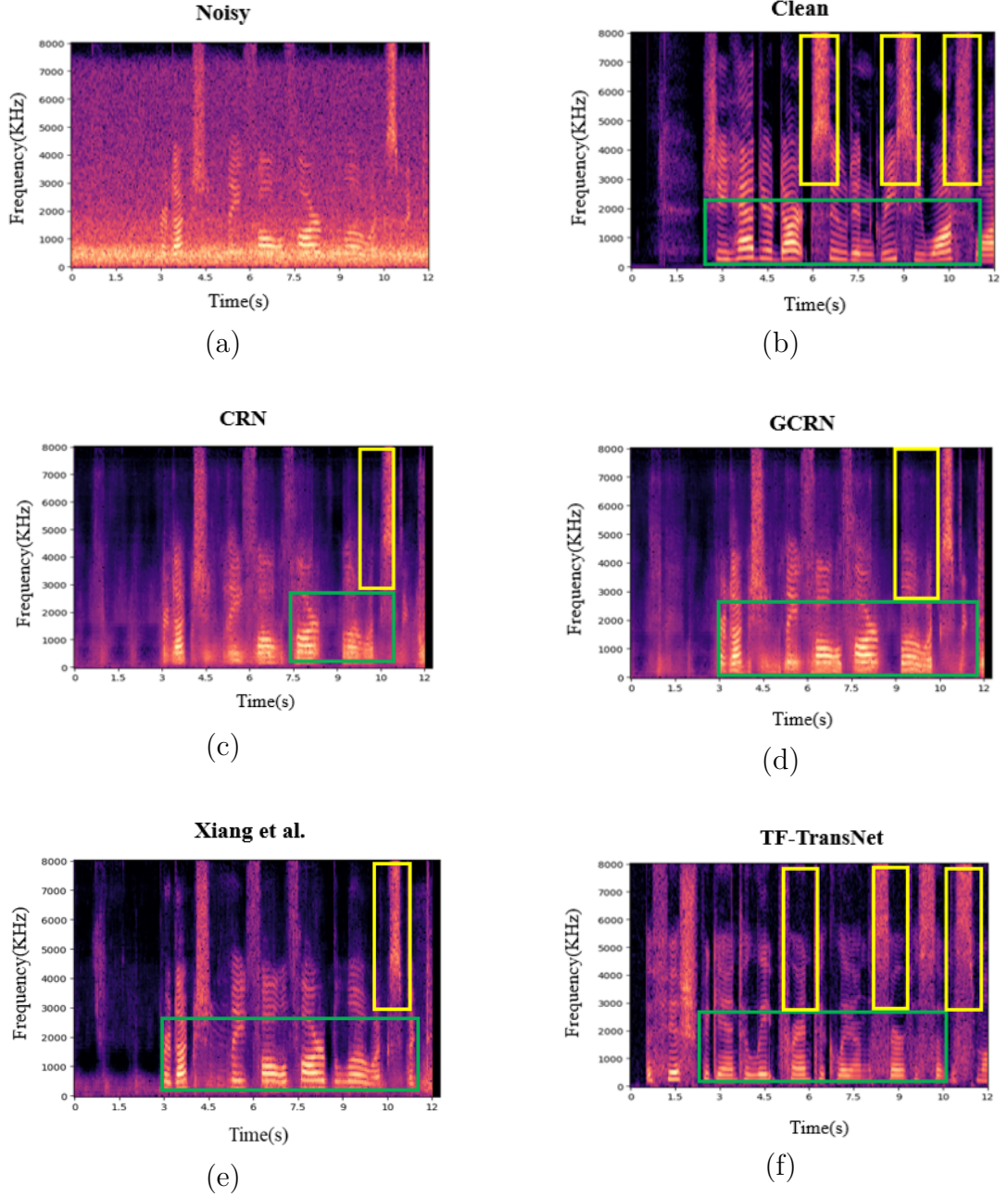


Figure 4.5: Spectrogram visualizations of (a) noisy and (b) clean speech samples, alongside the enhanced signal processed by existing SOTA models: (c) CRN, (d) GCRN, (e) Xiang et al. [50], and proposed (f) TF-TransNet.

with the differences clearly showing the areas of improvement.

Further analysis of the phase spectra provides additional insights into the TF-TransNet model's reconstruction process during speech enhancement. The Fig. 4.7 shows the phase spectra of noisy, clean, and enhanced speech signals, providing insights into the phase manipulation by the TF-TransNet model during speech enhancement. The noisy phase (top-left) shows a high degree of randomness due to noise, which disrupts the phase coherence. After processing, the enhanced phase (top-right) remains largely similar to the noisy phase. The

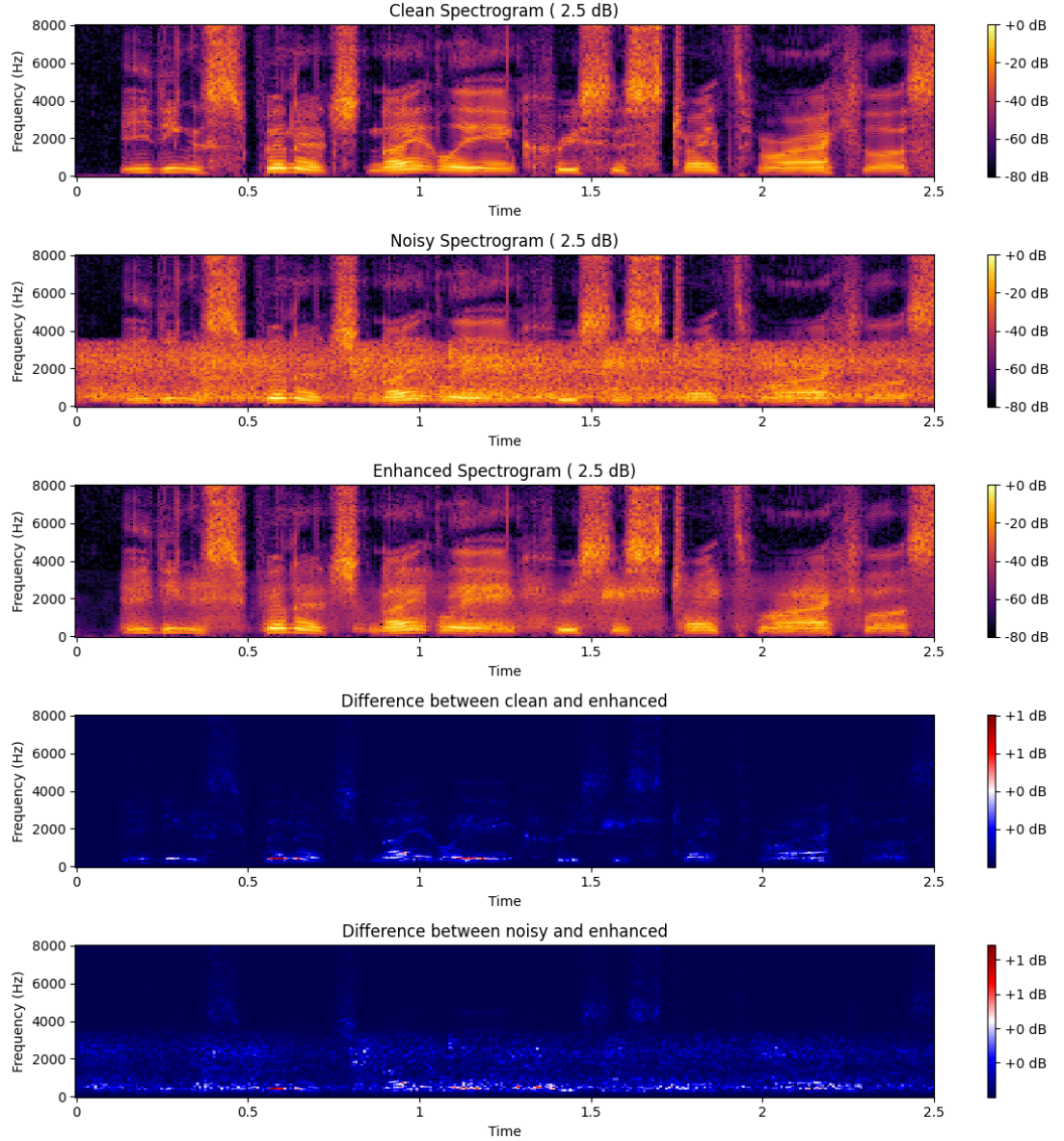


Figure 4.6: Spectrograms showing the difference between clean and enhanced, as well as noisy and enhanced speech signals at an SNR of 2.5 dB.

clean phase (bottom-left) displays a more structured and regular pattern, typical of natural speech. These observations highlight that the TF-TransNet model primarily uses the noisy phase in its reconstruction process, which results in relatively small phase discrepancies between the enhanced and clean signals. This approach underscores the model’s focus on preserving the original phase structure while improving the magnitude spectrum for noise suppression. The minimal phase difference between the clean and enhanced signals further suggests that the model achieves effective enhancement by focusing on magnitude-based features while maintaining phase coherence, leading to improved speech quality with minimal distortion.

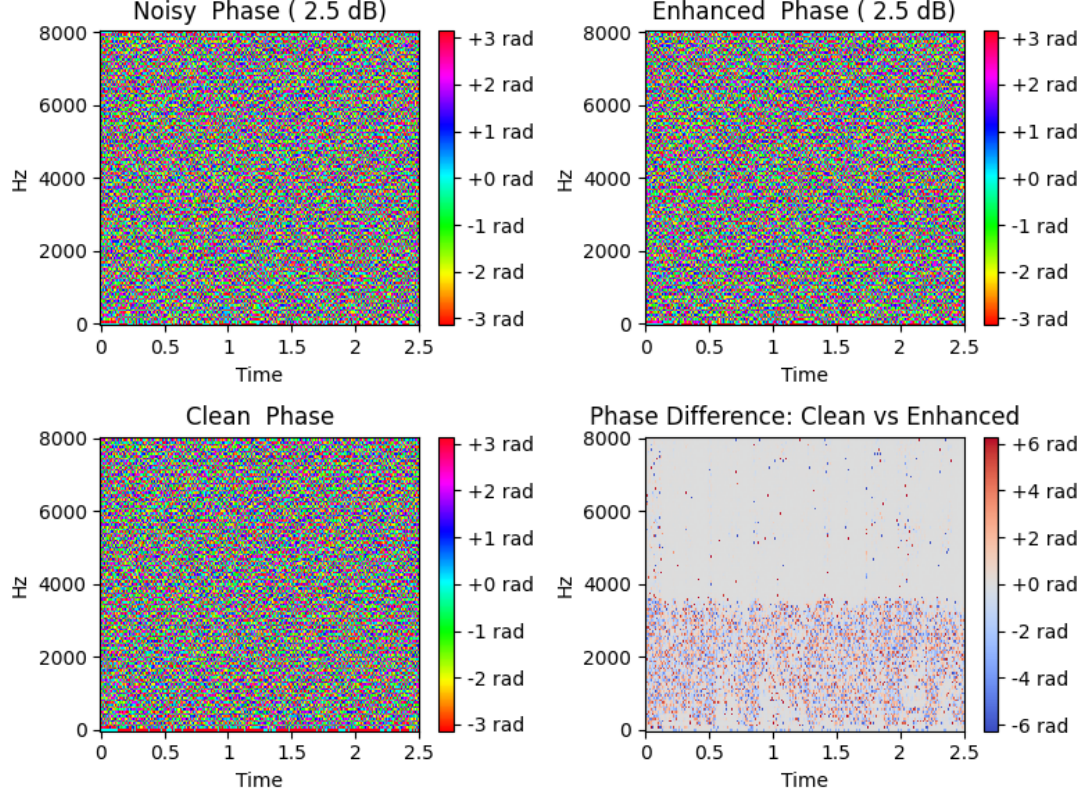


Figure 4.7: Unwrapped phase spectra of noisy, clean, and enhanced speech signals at an SNR of 2.5 dB, highlighting the TF-TransNet model’s reliance on the noisy phase for reconstruction and the minimal phase discrepancies between clean and enhanced signals.

4.4.4 Case Study

A case study is conducted in a practical setting for speech enhancement, where the subject recorded the utterance , "I am going to the store to buy some groceries," in a street environment with heavy traffic noise. In Fig.4.8 the first spectrogram illustrates the noisy speech signal, where background traffic noise predominates, rendering the speech difficult to discern. The second spectrogram represents clean, noise-free speech, characterized by distinct and intelligible frequency patterns. The third spectrogram displays the enhanced speech signal processed using the TF-TransNet model. While the model significantly reduces noise, complete elimination of the traffic noise is not achieved. Consequently, although significant level of noise suppression is evident, full separation of speech from background noise remains unattainable in this scenario.

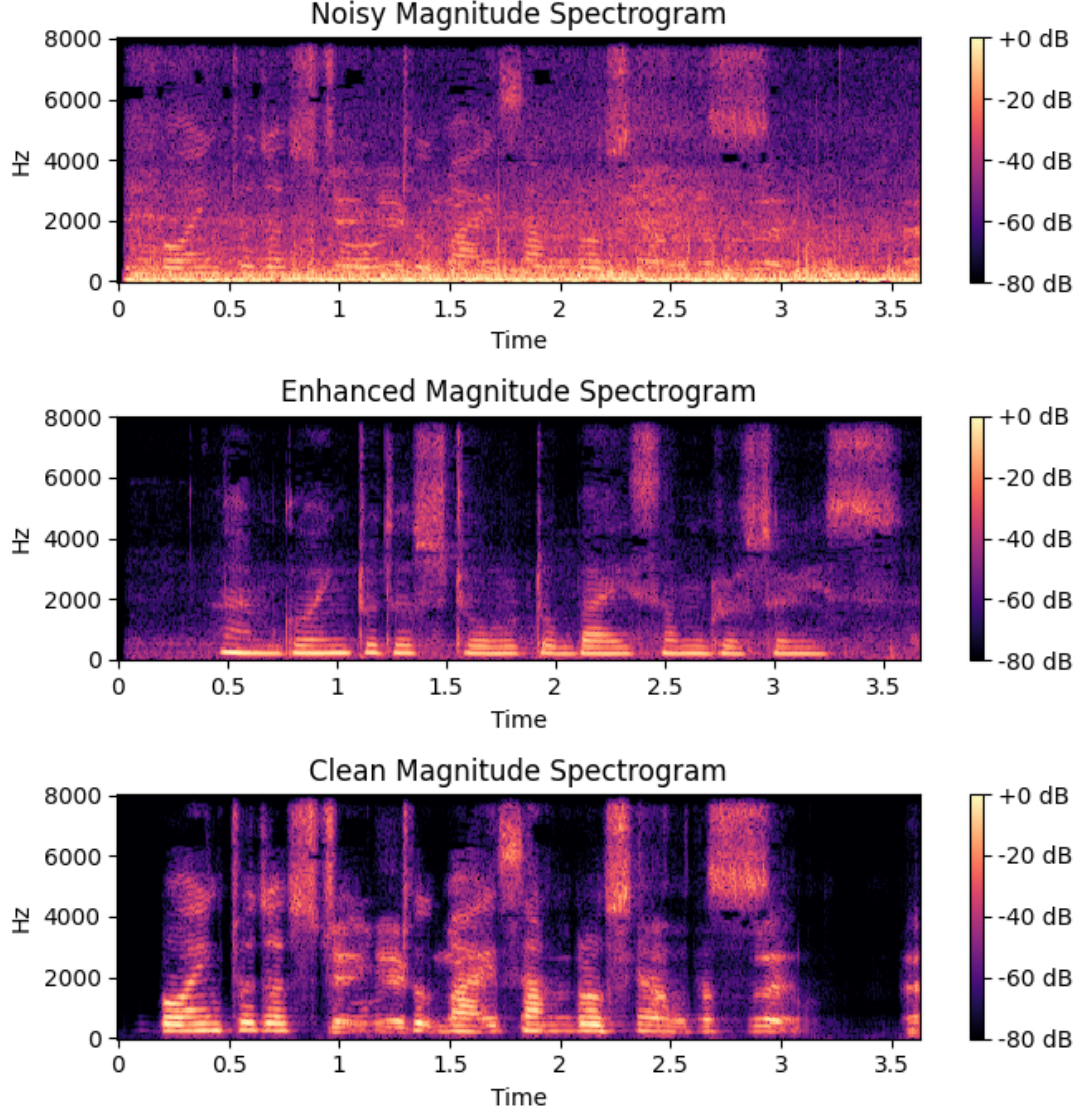


Figure 4.8: Spectrograms of the Noisy, Enhanced and Clean Speech Signals in a Practical Settings.

4.5 Summary

In conclusion, the TF-TransNet model integrates the TFA mechanism within the encoder, enabling it to selectively focus on critical time-frequency features of the speech signal, thereby enhancing clarity and intelligibility. Furthermore, the incorporation of the temporal-frequency (T-F) transformer in the bottleneck layer facilitates efficient information transfer from the encoder to the decoder, allowing the model to maintain contextual awareness and capture complex temporal and spectral dependencies. This combination significantly improves the model’s performance in suppressing noise and enhancing speech quality.

Chapter 5: Conclusion

This chapter provides a concise summary of the key findings of this research and highlights potential avenues for future work. Additionally, it includes a list of publications associated with this thesis.

5.1 Discussion

In conclusion, this work presents a masking-based strategy for speech enhancement that effectively addresses noise degradation in monaural speech. By employing the IRM within two distinct deep learning frameworks—U-Net and TF-TransNet, we have demonstrated significant improvements in speech intelligibility and quality. The U-Net model excels in moderate to high SNRs and familiar noise conditions, leveraging its encoder-decoder structure and skip connections to enhance information transfer. However, its limitations in low SNR scenarios and with unseen noise types highlight the need for further advancements. To overcome these challenges, the TF-TransNet introduces a TFA encoder and a T-F transformer layer, which enables the model to capture intricate temporal and spectral dependencies in the speech signal. The integration of multi-head self-attention and LSTM in T-F transformer layer further enhance the model’s ability to prioritize key features and maintain contextual awareness, resulting in superior performance compared to existing models like the CRN, GGCRN and masking-inpainting method. Overall, the findings of this research highlight the efficacy of the proposed approaches in improving speech enhancement across various noise conditions compared to mapping strategies.

5.2 Limitation

A notable limitation of the proposed speech enhancement approach is its dependence on the noisy phase for reconstructing the signal instead of improving the phase information itself. While the method focuses on enhancing the magnitude of the speech signal using techniques like the IRM, the phase remains unchanged. This reliance on the noisy phase can lead to the preservation of unwanted artifacts and distortions from the original signal, negatively impacting the

quality and clarity of the enhanced speech output. Since the phase is essential for accurately reconstructing how the speech sounds, its inadequacy can make the processed audio sound less clear and natural. Improving phase estimation techniques could greatly enhance the performance and usefulness of the speech enhancement system.

5.3 Future Scope

Future research in speech enhancement can focus on two important areas: Enhanced phase reconstruction and multichannel and binaural approaches. Enhanced phase reconstruction seeks to improve the quality of enhanced speech signals by developing methods to accurately estimate and reconstruct the noisy phase. This would tackle current limitations associated with phase reliance and minimize distortion, resulting in clearer and more intelligible speech. Furthermore, the incorporation of multichannel and binaural approaches can significantly enhance performance in complex acoustic environments. Utilizing multiple microphones or implementing binaural processing techniques allows for the capture of spatial information, which improves the model's ability to separate speech from background noise. Collectively, these advancements may greatly enhance the effectiveness and applicability of speech enhancement technologies in real-world scenarios.

5.4 List of Publications

The following publication is a direct consequence of the research carried out during the elaboration of the thesis and gives an idea of the progression that has been achieved.

- **Akter, K.**, Mamun, N., Hossain, M. A. (2023, February). A tf masking based monaural speech enhancement using u-net architecture. In 2023 International Conference on Electrical, Computer and Communication Engineering (ECCE) (pp. 1-5). IEEE.
- Namey, A. A., **Akter, K.**, Hossain, M. A., & Dewan, M. A. A. (2024). CochleaSpecNet: An attention based dual branch hybrid CNN-GRU network for speech emotion recognition using cochleagram and spectrogram. *IEEE Access*. (Accepted for publication).

- Orin, A. J., **Akter, K.**, Hossain, M. A., & Dewan, M. A. A. (2024). AD-FCINet: A fusion of convolution-involution network for Alzheimer’s disease staging with channel and spatial-specific features. *IEEE Transactions on Medical Imaging*. (Submitted for publication).

Bibliography

- [1] P. S. Siddala Vihari, A Sreenivasa Murthy and D. Naik, “Comparison of speech enhancement algorithms,” *Procedia computer science*, vol. 89, pp. 666–676, 2016.
- [2] . C. J. Wang, D., “Supervised speech separation based on deep learning: An overview,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702–1726, 2018.
- [3] N. Mamun, M. S. Zilany, J. H. Hansen, and E. E. Davies-Venn, “An intrusive method for estimating speech intelligibility from noisy and distorted signals,” *The Journal of the Acoustical Society of America*, vol. 150, no. 3, pp. 1762–1778, 2021.
- [4] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, “Perceptual evaluation of speech quality (pesq) - a new method for speech quality assessment of telephone networks and codecs,” in *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221)*, vol. 2, pp. 749–752, IEEE, May 2001.
- [5] J. H. Hansen and B. L. Pellom, “An effective quality evaluation protocol for speech enhancement algorithms,” in *Fifth International Conference on Spoken Language Processing*, 1998.
- [6] S. F. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27(2), pp. 113–120, 1979.
- [7] Y. Ephraim and H. L. V. Trees, “A signal subspace approach for speech enhancement,” *IEEE Transactions on speech and audio processing* 3, vol. no. 4, pp. 251–266, 1995.
- [8] P. Scalart, “Speech enhancement based on a priori signal to noise estimation,” *IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, vol. 2, pp. 629–632, 1996, May.
- [9] P. Smaragdis, “Convolutional speech bases and their application to supervised speech separation,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 1–12, 2006.
- [10] G. Hu and D. Wang, “Monaural speech segregation based on pitch tracking and amplitude modulation,” *IEEE Transactions on Neural Networks*, vol. 15(5), pp. 1135–1150., 2004.
- [11] Z. Jin and D. Wang, “A supervised learning approach to monaural segregation of reverberant speech,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17(4), pp. 625–638, 2009.

- [12] L. Y. H. Y. Kim, G. and P. C. Loizou, “An algorithm that improves speech intelligibility in noise for normal-hearing listeners.,” *The Journal of the Acoustical Society of America*, vol. 126(3), pp. 1486–1494, 2009.
- [13] K. Han and D. Wang, “A classification based approach to speech segregation,” *The Journal of the Acoustical Society of America*, vol. 132(5), pp. 3475–3483, 2012.
- [14] D. J. D. L.-R. Xu, Y. and C.-H. Lee, “A regression approach to speech enhancement based on deep neural networks,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23(1), pp. 7–19, 2015.
- [15] W. W. Yu, Y. and P. Han, “Localization based stereo speech source separation using probabilistic time-frequency masking and deep neural networks,” *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 7(1), pp. 1–18, 2016.
- [16] J. Chen and D. Wang, “Dnn based mask estimation for supervised speech separation,” *In Audio Source Separation*, vol. 7(1), pp. 207–235, 2018.
- [17] M. A. K. Karjol, Pavan and P. K. Ghosh, “Speech enhancement using multiple deep neural networks,” *In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. pp, pp. 5049–5052, 2018.
- [18] Y. Xu, J. Du, Z. Huang, L. R. Dai, and C. H. Lee, “Multi-objective learning and mask-based post-processing for deep neural network based speech enhancement,” *arXiv preprint arXiv:1703.07172*, 2017.
- [19] T. Gao, J. Du, L. R. Dai, and C. H. Lee, “Snr-based progressive learning of deep neural network for speech enhancement,” in *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech)*, 2016.
- [20] B. Y. Xia and C. C. Bao, “Speech enhancement with weighted denoising autoencoder,” in *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech)*, pp. 3444–3448, August 2013.
- [21] A. Maas, Q. V. Le, T. M. O’Neil, O. Vinyals, P. Nguyen, and A. Y. Ng, “Recurrent neural networks for noise reduction in robust asr,” in *ISCA Interspeech*, pp. 22–25, 2012.
- [22] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, “Investigating rnn-based speech enhancement methods for noise-robust text-to-speech,” in *ISCA Speech Synthesis Workshop*, pp. 146–152, 2016.
- [23] F. Weninger, J. R. Hershey, J. L. Roux, and B. Schuller, “Discriminatively trained recurrent neural networks for single-channel speech separation,” in *2014 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pp. 577–581, 2014.
- [24] J. Chen and D. Wang, “Long short-term memory for speaker generalization in supervised speech separation,” *The Journal of the Acoustical Society of America*, vol. 141, no. 6, pp. 4705–4714, 2017.
- [25] M. Wollmer, Z. Zhang, F. Weninger, B. Schuller, and G. Rigoll, “Feature enhancement by bidirectional lstm networks for conversational speech recognition in highly non-stationary noise,” in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6822–6826, 2013.

- [26] T. Kounovsky and J. Malek, “Single channel speech enhancement using convolutional neural network,” in *2017 IEEE International Workshop of Electronics, Control, Measurement, Signals and their Application to Mechatronics (ECMSM)*, p. 15, IEEE, May 2017.
- [27] D. Rethage, J. Pons, and X. Serra, “A wavenet for speech denoising,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2018.
- [28] K. Tan and D. Wang, “A convolutional recurrent neural network for real-time speech enhancement,” in *ISCA Interspeech*, pp. 3229–3233, 2018.
- [29] X. Hao, C. Shan, Y. Xu, S. Sun, and L. Xie, “An attention-based neural network approach for single channel speech enhancement,” in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6895–6899, IEEE, 2019.
- [30] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” in *International Conference on Learning Representations (ICLR)*, 2015.
- [31] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, and Y. Bengio, “End-to-end attention-based large vocabulary speech recognition,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016.
- [32] C. Shan, J. Zhang, Y. Wang, and L. Xie, “Attention-based end-to-end speech recognition on voice search,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.
- [33] C. Shan, J. Zhang, Y. Wang, and L. Xie, “Attention-based end-to-end models for small-footprint keyword spotting,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.
- [34] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Advances in Neural Information Processing Systems*, pp. 5998–6008, 2017.
- [35] W. Yu, J. Zhou, H. Wang, and L. Tao, “Setransformer: Speech enhancement transformer,” *Cognitive Computation*, pp. 1–7, 2022.
- [36] J. Chen, Q. Mao, and D. Liu, “Dual-path transformer network: Direct context-aware modeling for end-to-end monaural speech separation,” *arXiv preprint arXiv:2007.13975*, 2020.
- [37] A. Narayanan and D. Wang, “Ideal ratio mask estimation using deep neural networks for robust speech recognition,” in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 7092–7096, IEEE, 2013.
- [38] D. S. Williamson, Y. Wang, and D. L. Wang, “Complex ratio masking for monaural speech separation,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 3, pp. 483–492, 2016.
- [39] Y. Wang, A. Narayanan, and D. Wang, “On training targets for supervised speech separation,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1849–1858, 2014.

- [40] D. Wang, “On ideal binary mask as the computational goal of auditory scene analysis,” in *Speech Separation by Humans and Machines*, pp. 181–197, Boston, MA: Springer, 2005.
- [41] Y. Wang, A. Narayanan, and D. Wang, “On training targets for supervised speech separation,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1849–1858, 2014.
- [42] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, “Phase sensitive and recognition-boosted speech separation using deep recurrent neural networks,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, pp. 708–712, IEEE, 2015.
- [43] K. Tan and D. Wang, “Learning complex spectral mapping with gated convolutional recurrent networks for monaural speech enhancement,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 380–390, 2019.
- [44] D. S. Williamson, Y. Wang, and D. L. Wang, “Complex ratio masking for monaural speech separation,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 3, pp. 483–492, 2016.
- [45] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III*, vol. 18, pp. 234–241, Springer International Publishing, 2015.
- [46] V. Zue, S. Seneff, and J. Glass, “Speech database development at mit: Timit and beyond,” *Speech Communication*, vol. 9, no. 4, pp. 351–356, 1990.
- [47] N. Mamun, S. Majumder, and K. Akter, “A self supervised convolutional neural network approach for speech enhancement,” in *2021 5th International Conference on Electrical Engineering and Information & Communication Technology (ICEE-ICT)*, pp. 1–5, IEEE, 2021.
- [48] K. Wang, B. He, and W.-P. Zhu, “Tstnn: Two-stage transformer based neural network for speech enhancement in the time domain,” in *Proc. ICASSP*, pp. 7098–7102, 2021.
- [49] S. Lv, Y. Jv, and L. Xie, “Uformer: A u-net based dilated complex & real dual-path conformer network for simultaneous speech enhancement and dereverberation,” in *Proc. ICASSP*, pp. 7417–7421, 2022.
- [50] Y.-X. Lu, Y. Ai, and Z.-H. Ling, “Explicit estimation of magnitude and phase spectra in parallel for high-quality speech enhancement,” *arXiv preprint arXiv:2308.08926*, 2023.
- [51] J. Chen, Q. Mao, and D. Liu, “Dual-path transformer network: Direct context-aware modeling for end-to-end monaural speech separation,” *arXiv preprint arXiv:2007.13975*, 2020.
- [52] K. Tan, J. Chen, and D. Wang, “Gated residual networks with dilated convolutions for monaural speech enhancement,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 1, pp. 189–198, 2018.

- [53] X. Hao, X. Su, S. Wen, Z. Wang, Y. Pan, F. Bao, and W. Chen, “Masking and inpainting: A two-stage speech enhancement approach for low snr and non-stationary noise,” in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 6959–6963, IEEE, May 2020.