

Master of Science in Electronics and Telecommunication Engineering



Deep Learning based Bandwidth Extension and Speech Restoration

by

Taieba Taher

ID: 19METE011P

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.

May, 2024

CERTIFICATION

The thesis titled **Deep Learning based Bandwidth Extension and Speech Restoration** submitted by **Taieba Taher**, Roll No. **19METE011P**, 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 **May 2024**.

BOARD OF EXAMINERS

1. _____
Dr. Md Azad Hossain Chairman (Supervisor)
Professor
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

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

Taieba Taher
ID: 19METE011P

Author's 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 can be uploaded to the repository of the library, Chittagong University of Engineering and Technology, 1 year from the day of submission.

I authorize Chittagong University of Engineering and Technology to lend 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 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

Taieba Taher

ID: 19METE011P

Department of Electronics and Telecommunication Engineering (ETE),
Chittagong University of Engineering and Technology (CUET).

APPROVAL by the SUPERVISOR

This is to certify that TAIEBA TAHER 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 CUET's PLAGIARISM and ACADEMIC INTEGRITY regulations.

Signature of the Supervisor

Dr. Md. Azad Hossain
Professor
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.

1. **T. Taher**, N. Mamun, and M. A. Hossain, “A joint bandwidth expansion and speech enhancement approach using deep neural network,” in 2023 International Conference on Electrical, Computer and Communication Engineering (ECCE), pp. 1–4, IEEE, 2023

This thesis dedicated to,
my respected parents,

Md. Abu Taher and Fatema Khatun

and my beloved spouse,

Lt. Cdr. Tauhid

for their love, endless support, encouragement, and sacrifices.

Acknowledgment

First and foremost, I express my gratitude to Almighty Allah for granting me the strength to complete this thesis. The sense of fulfillment derived from its accomplishment would be inadequate without acknowledging those whose unwavering support and guidance have been instrumental in its fruition. I extend my heartfelt thanks to several individuals, without whom this endeavor would not have been feasible. I extend my sincere appreciation to Professor Dr. Md Azad Hossain for his unwavering support and guidance as my supervisor. Additionally, I am indebted to Dr. Nursadul Mamun for his mentorship throughout my thesis journey, providing invaluable encouragement and fostering a positive environment. His dedication to research in Speech Signal Processing, boundless patience, and willingness to nurture my growth have significantly contributed to my development in this field. I am profoundly grateful to both individuals for their continuous encouragement and insightful contributions.

Finally, I am thankful to all the members of faculty members, and staff of the department of ETE, CUET, for their support. My family deserves special thanks for their continuing love, support, encouragement, and sacrifices.

Abstract

Speech Bandwidth Extension (SBE) remains a challenging task in speech processing. It involves the intricate process of estimating missing frequency components in low-resolution signals to reconstruct high-resolution transformations. SBE is a technique used to enhance the quality of speech signals by expanding the frequency range of the audio. This endeavor is fundamental for enhancing speech quality, naturalness, and intelligibility, particularly in scenarios marred by background noise and channel distortions. Low-resolution signals often lack high-frequency components due to limited sampling rates and band-limiting, as dictated by the Nyquist-Shannon Sampling Theorem. The inherent loss of high-frequency details stems from aliasing, quantization errors, and insufficient granularity in sampling, which poses significant challenges in restoring full-band speech. Addressing these issues is critical to advancing SBE techniques. Traditional methodologies heavily relied on statistical algorithms and learning techniques such as Multi-Layer Perceptrons (MLP). However, the paradigm shift brought about by Deep Learning has revolutionized this domain, opening avenues for more efficient and effective solutions. This thesis delves into deep learning-based approaches for addressing the complexities of BWE and speech restoration, particularly emphasizing noisy environments. This research explores three distinct augmentations of a frequency-domain deep learning network, each tailored to tackle specific challenges encountered in bandwidth extension and speech enhancement. The first approach introduces a joint bandwidth expansion and speech enhancement paradigm utilizing Deep Neural Networks (DNNs). This approach is meticulously designed to simultaneously expand the bandwidth of speech signals, reduce noise interference, and maintain the quality and intelligibility of the speech. Leveraging the inherent capabilities of DNNs, this methodology accurately estimates missing speech components. It characterizes noise profiles within degraded signals, synthesizing high-fidelity full-band speech from limited bandwidth inputs. The experimentation demonstrates the superior performance of this DNN-based approach, surpassing conventional methods and presenting promising avenues for real-world applications. Stepping beyond conventional methodologies, the second approach introduces an end-to-end frequency-domain framework, aptly named the Robust extension-plus-enhancement of speech utilizing Dual-former Network (RDNet). RDNet represents a paradigm shift in BWE, aiming to recover full-band speech from noisy low-band signals directly. By intricately integrating speech enhancement and ideal bandwidth extension modules

within a unified framework, RDNet demonstrates remarkable performance across diverse noisy environments. Leveraging short-time Fourier transform (STFT) features for enhancement, RDNet achieves significant improvements in critical metrics such as Short-Time Objective Intelligibility (STOI), Perceptual Evaluation of Speech Quality (PESQ), and log-spectral distortion (LSD), underscoring its efficacy and potential for practical deployment. Finally, this study’s investigation culminates in introducing the conformer-motivated super-denoised network (CSDNet), a novel approach tailored to mitigate the mismatch problem inherent in domain-specific encoders/decoders. By embedding a conv Unet with T-F transformation layers, CSDNet reduces dependency on training data and outperforms recent baseline methods across various objective metrics. Moreover, CSDNet exhibits favorable subjective performance in comparative studies, reaffirming its suitability for real-time applications. Through these meticulously crafted approaches, this thesis aims to advance the frontier of deep learning-based speech bandwidth extension and speech restoration. By addressing the core challenges of high-frequency component loss in low-resolution signals and proposing innovative solutions, this work offers insights and methodologies, proposes practical solutions for tackling real-life problems, and contributes to the wider conversation in the fields of speech processing and audio engineering.

Keywords: Speech Bandwidth Extension, Speech Enhancement, Speech Restoration, Deep learning network, Speech intelligibility and quality, Speech communication, Deep Neural Network, Unet, Conformer, Transformer, Robust bandwidth extension.

Contents

Acknowledgment	i
Abstract	ii
1 Introduction	1
1.1 Research Background	1
1.2 Problem Statement	3
1.3 Aims and Objectives	4
1.4 Significance and Scope of study	5
1.5 Organization of Thesis	6
2 Literature Review	7
2.1 Production and Transmission of Speech	7
2.2 Impact of Resolution Limitations on Speech Bandwidth Extension and Restoration	8
2.2.1 Sampling Rate and the Nyquist-Shannon Theorem	8
2.2.2 Aliasing and Its Effects	8
2.2.3 Quantization Resolution and High-Frequency Sensitivity	9
2.2.4 Band-Limiting Filters in Signal Acquisition	9
2.2.5 Granularity and Information Loss	9
2.3 Speech Bandwidth Extension Algorithms	10
2.3.1 Spectral Bands Mirroring	11
2.3.2 Approaches Utilizing Machine Learning	11
2.3.3 Approaches Utilizing Deep Learning	12
2.4 Research Gap	14
3 Joint Speech Bandwidth Expansion and Restoration	16
3.1 Motivation	16
3.2 Proposed Architecture	17
3.2.1 Model Description	18
3.2.2 Reconstruction of Signal	19
3.3 Experiments and Results	20
3.3.1 Dataset Description	20
3.3.2 Experimental Setup	20
3.3.3 Results	21
3.4 Summary	24

4	RDNet: Robust extension-plus-enhancement of speech utilizing Dual-former Network	26
4.1	Motivation	26
4.2	Proposed Architecture	28
4.2.1	Model Description	29
4.2.2	Dual-former	30
4.3	Experiments and Results	33
4.3.1	Dataset Description	33
4.3.2	Experimental setup	34
4.3.3	Results	35
4.4	Summary	37
5	CSDNet:Conformer Motivated Super Denoised Network	39
5.1	Motivation	39
5.2	Proposed Architecture	41
5.2.1	Encoder-Decoder	42
5.2.2	T-F Transformation Layer	43
5.2.3	Parallel Recurrent Neural Network (PRNN)	45
5.2.4	Conformer Block	46
5.2.5	Skip Connection	46
5.3	Experimental Setup	47
5.3.1	Speech Corpus	47
5.3.2	Experimental Framework	48
5.3.3	Evaluation Metrics	48
5.3.4	Baseline Networks	49
5.4	Results	50
5.4.1	Assessment Category	50
5.4.2	Comparative Objective Evaluation of Speech Intelligibility and Quality Metrics	51
5.4.3	Evaluation of CSDNet Performance Across Various Noise Conditions and SNRs	52
5.4.4	Impact of Phase on Reconstructed Speech	52
5.4.5	Investigating the Component Effectiveness in Proposed Model through Ablation Studies	55
5.4.6	Comparative Analysis	57
5.4.7	Comparative Subjective Evaluation of Speech Intelligibility and Quality Metrics	58
5.5	Summary	59
6	Conclusion	61
6.1	Future Scope	62
6.2	List of Publications	64

List of Figures

3.1	Overall block diagram of proposed DNN-based joint speech bandwidth expansion and speech restoration.	18
3.2	Spectrogram of one test utterance with SNR 10 dB, (a) Noisy low band signal spectrogram, (b) Predicted full band signal spectrogram, (c) Original full band signal spectrogram.	24
4.1	The process involves two key steps: First, the noisy narrowband signal undergoes enhancement to eliminate noise (Step 1). Second, the improved narrowband signal is extended in bandwidth to produce the clean fullband signal(Step 2) [1].	27
4.2	An illustration of the proposed network architecture(RDNet). The processing chain depicted in (a) illustrates the comprehensive prediction core, utilizing the Short-Time Fourier Transform (STFT) signal transformation, high-resolution spectrogram prediction, and the Inverse STFT process. Meanwhile, the modular components illustrated in (b) represent individual cores, each equipped with learned feature transformations based on 2D-Conv encoder-decoder layers, and a dual-former layer as bottleneck.	28
4.3	Architecture of Dual-former block consisting of a transformer, a conformer, frequency transformation and instant layer normalization layers.	30
4.4	An illustration of (a) transformer and (b) conformer layer used in Dual-former block.	31
4.5	Spectrogram of one test utterance with SNR 5 dB, (a) Noisy low band signal spectrogram, (b) Predicted full band signal spectrogram, (c) Clean full band signal spectrogram with oracle phase.	36
5.1	An illustration of the proposed network architecture(CSDNet). The processing chain depicted in (a) illustrates the comprehensive prediction core, utilizing the Short-Time Fourier Transform (STFT) signal transformation, high-resolution spectrogram prediction, and the Inverse STFT process. Meanwhile, the modular components illustrated in (b) represent individual cores, each equipped with learned feature transformations based on 2D-Conv encoder-decoder layers, a Time-Frequency (T-F) transformation layer, a Parallel Recurrent Neural Network (PRNN), and a Conformer module.	42
5.2	Architecture of T-F Transformation Layer.	44

5.3	Architecture of Parallel RNN(PRNN) visualizing the data dividing into segments, analyzing those segments, and finally reassembling them.	45
5.4	Objective measurements on various phase of input test sets of proposed models. Oracle, Imaged__noisy, Imaged__inverse and Griffin Lim, respectively are four different phases for the reconstruction of test sets. Orange, Yellow, and Green bars represent the SNRs of -5 dB, 0dB and 5 dB, respectively.	53
5.5	Qualitative results. Spectrogram of (a) clean full band speech, (b) noisy low band speech, (c) DNN output, (d) CNN output, (e) TFiLM-Unet output, (f) AFiLM-Unet output and (g) CSDNet output (Proposed).	56
5.6	Subjective Speech Enhancement Performance based on (a) Speech Intelligibility, (b) Speech Quality	59

List of Tables

2.1	Brief literature summary on a few recent works of deep learning-based speech bandwidth extension models concerning the domain, features, models, Contributions, and their limitations.	15
3.1	Effects of different phases in Clean Scenario.	23
3.2	Performance Evaluation of Full Band Noise (SSN)	23
3.3	Performance Evaluation of Low Band Noise (Car)	23
4.1	Evaluation of objective scores in seen SNRs utilizing the TIMIT and DNS challenge dataset.	35
4.2	Evaluation of objective scores in Unseen SNRs utilizing the TIMIT and DNS challenge dataset.	36
5.1	Results of the metrics for assessing intelligibility and quality of speech are presented in this table, including the training condition. Note that all results are evaluated under the noisy TIMIT testset. Scores were calculated for STOI(0 ~ 1), SOPM(0 ~ 1), PESQ(-0.5 ~ 4.5), CSIG, CBAK, COVL are respectively (1 ~ 5).	51
5.2	Quantitative evaluation of proposed Speech Bandwidth Extension algorithm (CSDNet) for six SNR groups in both seen and unseen noises.	52
5.3	Evaluation of Model Components through Ablation Study.	55
5.4	A Comparative result of all metrics among different existing models is presented in this table. The upper arrow represents that higher is better, and vice versa. Note that all results are evaluated under both clean and noisy TIMIT test sets. The approaches are performed in both frequency domain and time domain.	57

Chapter 1

Introduction

The chapter will highlight the significance of human speech in communication, emphasizing its complex interplay of linguistic content and emotional nuances. It discusses challenges posed by gender-based differences in voice frequencies and the limitations of traditional telephony systems in preserving speech quality due to bandwidth constraints. The focus shifts to bandwidth extension methods aimed at enhancing high-frequency content perception in speech signals, with a particular emphasis on speech restoration techniques. The concept of speech restoration, also known as speech enhancement, is introduced as a critical component in improving speech quality by reducing distortions and noise. Bandwidth extension methods, including perceptive and imperceptive strategies, are discussed, along with their applications in speech restoration efforts. The introduction underscores the potential of speech restoration through artificial bandwidth extension in various communication scenarios, from telephone communication to historical recordings, to enrich the overall communication experience.

1.1 Research Background

Human speech stands as a cornerstone of communication, weaving linguistic content with emotional nuances conveyed through tone and cadence. It manifests in voiced pulses carrying periodic information and unvoiced noise, each laden with significance. Within this intricate tapestry of communication, pitch emerges as a pivotal element, serving as a statistical measure of temporal variations that adds depth to conversational speech. Notably, gender differences are reflected in the fundamental frequencies of male and female voices, with male voices typically exhibiting mean and standard deviation values on a logarithmic frequency scale approximately half that of females. This disparity poses challenges in scenarios where speech transmission occurs through systems with limited bandwidth, such

as traditional telephony systems. These systems, constrained by financial considerations, prioritize clear communication at the expense of reduced speech quality. As society continues to pursue optimal sound quality, the quest for efficiency persists. Bandwidth extension emerges as a solution, wherein speech signals undergo segmentation, processing, and recombination to mitigate the limitations of restricted bandwidth transmission systems. However, achieving seamless integration of processed segments to maintain perceptual coherence poses a significant challenge. Several methods of bandwidth extension have been introduced, classified based on factors such as the spectral band's frequency range and the location of signal bandwidth extension. This study focuses on extending bandwidth to perceive high-frequency content in reproduced conversational speech signals. For instance, in contemporary audio storage and transmission formats like MPEG1 layer-3 (MP3), perceptual coding techniques at high compression ratios introduce distortions that affect high-frequency components, leading to diminished sound quality. Speech restoration, also known as speech enhancement, emerges as a critical component in this context. It aims to improve the quality of speech signals degraded during transmission or recording. By employing advanced signal processing techniques, speech restoration algorithms enhance speech clarity, reduce noise, and reconstruct missing or distorted components, thereby enhancing overall intelligibility and naturalness. Bandwidth extension methods, whether perceptive or imperceptive[2], play a vital role in speech restoration efforts. Perceptive strategies aim to reproduce high-quality full-band signals by extrapolating missing information, and leveraging correlations between lower and higher frequency bands. On the other hand, imperceptive strategies employ artificial bandwidth extension (BWE) techniques to analyze and augment restricted-range speech signals, producing approximated wideband speech signals for playback. The implementation of BWE in speech transmission involves filtering the full-band microphone signal before digital conversion and transmission across the telephone network. At the receiving end, the narrowband signal undergoes analysis and augmentation, culminating in the generation of an approximated wideband speech signal for playback. This technique, pioneered by Croll[3], has shown promise in enhancing audio fidelity in broadcast programs and historical speech recordings. Speech restoration through BWE holds immense potential in various applications, from improving telephone communication to enhancing historical recordings. By leveraging advancements in signal processing and utilizing supplementary wideband recordings when available, speech restoration algorithms can effectively enhance speech clarity and fidelity, enriching the overall communication experience.

1.2 Problem Statement

In challenging acoustic environments where multiple speakers or background noises coexist, humans can selectively focus their attention on a particular speaker’s words while ignoring distractions. This ability forms the basis of speech restoration, also known as speech enhancement, algorithms, which aim to mimic this selective attention process by masking out background noise and enhancing important speech content. Speech restoration algorithms find wide application as pre-processing modules in various real-world scenarios, including automatic speech recognition (ASR), speaker identification, and the design of hearing aids[4][5]. Over several decades, researchers have proposed various statistical approaches to simulate human selective attention, such as subspace algorithms[6], minimum-mean-square-error (MMSE)-based spectral amplitude estimators[7], and soft-decision noise suppression filters[8] [7][9]. These algorithms strive to estimate optimal multiplicative masks through statistical inference to attenuate background noises. The emergence of deep neural networks (DNNs) has led to learning-based models for speech restoration, such as feed-forward networks[10], recurrent and prolonged short-term memory (LSTM) networks[5], and others. However, frequency-domain approaches have shown limitations, including suboptimal feature extraction using short-time Fourier transform (STFT) and challenges in phase reconstruction, leading to compromised speech quality[5]. While post-processing methods have been proposed to address phase estimation issues, their effectiveness remains limited[5]. Recent advancements have introduced end-to-end frameworks that operate directly on noisy waveforms in the time domain, bypassing the need for explicit short-time Fourier transform (STFT) operations. These time-domain approaches, such as fully convolutional networks (FCNs)[11], Wave-U-Net architectures[12], and Conv-Tasnet[13], offer promising solutions by optimizing encoder-decoder structures and eliminating phase estimation challenges. Despite these advancements, learning-based speech bandwidth extension (SBE) frameworks need help in real-life scenarios due to mismatches between training and testing data distributions. Factors contributing to this mismatch include unseen speakers, accents, noises, channel distortions, and sensitive time-domain encoder/decoder operations[14]. This thesis will discuss a few techniques that can be used to address the prior problems that arise from the following three factors-

- to combine speech bandwidth extension and speech enhancement in joint training neural networks to mitigate mismatches caused by channel distortions.

- to adapt speech enhancement models trained on source domain data to target domain data, effectively reducing mismatches caused by unseen noises in the test data for speech restoration.
- to combine directly frequency-domain and time-domain networks to achieve balanced results, although the performance of such methods still needs to improve.

The proposed thesis aims to build upon the insights and limitations identified in prior studies to develop novel approaches for addressing speech bandwidth extension and restoration challenges. These approaches will consider unseen noises, channel distortions, and sensitive time-domain encoder/decoder operations to advance state-of-the-art speech enhancement algorithms.

1.3 Aims and Objectives

The focus of this thesis is to investigate and advance the state-of-the-art speech bandwidth extension (SBE) and speech restoration (SR) techniques, specifically leveraging generative diffusion models for enhancing speech quality in various real-world scenarios. The major objectives of this work are-

- To delve into speech bandwidth extension (SBE) and speech restoration (SR) to improve speech quality. This technique aims to increase the waveform resolution of speech signals, particularly in scenarios where high-frequency components are lacking or degraded, thereby improving the intelligibility and quality of speech in real-world applications.
- To investigate the effectiveness of discriminative models in speech restoration tasks. It will compare their performance against generative diffusion models using the deep learning architecture, ensuring a fair and comprehensive evaluation of the models.
- To assess the performance of different discriminative models across different types of speech corruptions, including background noise, interfering speakers, room acoustics, and channel degradation.

This thesis aims to advance speech processing techniques by achieving the aforementioned goals and introducing new perspectives and methods to speech restoration and bandwidth extension.

1.4 Significance and Scope of study

Speech Bandwidth Extension (SBE) is a pioneering force driving advancements in audio technology, offering a gateway to transcend traditional audio boundaries and ushering in a realm of more prosperous and more immersive auditory experiences. Its paramount relevance lies in its capacity to enhance audio quality and expand the frequency range of audio signals, thereby enabling new avenues for artistic expression. The scope of this study is comprehensive, addressing the overarching challenge of reconstructing high-frequency signals within the domain of Speech Bandwidth Extension (SBE). It encompasses algorithmic development, audio quality enhancement, real-time processing capabilities, noise reduction strategies, accessibility considerations, interdisciplinary collaboration, and future-proofing of SBE solutions. This thorough approach aims to delve into the nuances of high-frequency signal reconstruction, instilling confidence in the efficacy of our research. A focal point of future exploration within Speech Bandwidth Extension (SBE) is mitigating spectrum discontinuity issues inherent in high-frequency signal reconstruction. This endeavor necessitates the development of sophisticated algorithms coupled with the integration of machine learning methodologies, contextual analysis frameworks, perceptual quality evaluation metrics, cross-modal applications, and enhancements in accessibility. The overarching objectives of this endeavor encompass the elevation of audio quality standards and the facilitation of interoperability within rapidly evolving technological landscapes. Furthermore, within the realm of Speech Bandwidth Extension (SBE), there lies a substantial opportunity for advancements in high-frequency signal reconstruction techniques. This entails an in-depth investigation into adaptive techniques capable of dynamically adjusting the extended bandwidth by available resources and network conditions. By optimizing the reconstruction of high-frequency signals across diverse settings, such adaptive techniques hold the potential to ensure effective and high-quality audio transmission. Moreover, the study explores real-time adaptation techniques to facilitate on-the-fly high-frequency signal reconstruction. These real-time adaptation techniques are integral components in the seamless integration of Speech Bandwidth Extension (SBE) solutions by enabling optimal audio quality and efficient bandwidth utilization in dynamic communication scenarios.

1.5 Organization of Thesis

This thesis is structured into six chapters, each addressing distinct aspects of speech bandwidth extension and enhancement:

Section 1 serves as an introduction, outlining the motivation behind the research, delineating its contributions, and providing a succinct overview of the thesis.

Section 2 delves into the foundational concepts of speech bandwidth extension, elucidating the intricacies involved. It proceeds to review contemporary techniques in the field, shedding light on the prevalent challenge of mismatched probability distributions between training and test data.

Section 3 introduces a novel paradigm for joint bandwidth expansion and speech enhancement, leveraging Deep Neural Networks (DNNs). This chapter meticulously details the methodology, encompassing the augmentation of speech bandwidth, noise mitigation, and preservation of speech quality and intelligibility. Additionally, it elucidates the experimental setup and evaluation framework.

In Section 4, the focus shifts towards an end-to-end frequency-domain framework termed the Robust extension-plus-enhancement of speech utilizing Dual-former Network (RDNet). This chapter critically assesses the limitations of prior approaches and proposes a comprehensive solution for directly recovering full-band speech from noisy low-band signals. It proceeds to elaborate on the experimental setup and provides in-depth analysis of the experimental results.

Section 5 introduces the Conformer motivated Super Denoised network (CS-DNet), a novel approach aimed at mitigating the mismatch problem inherent in domain-specific encoders/decoders. Beginning with an exposition of the constraints of time-domain approaches, this chapter unveils the proposed methodology and delves into the experimental setup and results analysis.

Section 6 draws the thesis to a close, summarizing its contributions and offering insights into potential future directions for research in the field.

Chapter 2

Literature Review

This chapter provides a concise overview of speech production, transmission, and representation, emphasizing their significance in understanding Speech Bandwidth Extension (SBE) algorithms. It discusses various representations used in speech signal analysis and processing, including spectral, temporal, statistical, pitch-related, contextual, and hybrid representations. Additionally, the chapter explores the challenges and methodologies of SBE algorithms, categorizing them into time domain and frequency domain methods. It highlights the evolution of SBE techniques, including Spectral band mirroring (SBM) and approaches utilizing Machine Learning and Deep Learning. Overall, the chapter lays the foundation for understanding the subsequent sections of the study by summarizing key concepts and recent advancements in speech production, transmission, representation, and SBE algorithms.

Section 2.1 provides insight into the production and transmission of speech. Section 2.2 outlines the impact of Resolution Limitations on Speech Bandwidth Extension and Restoration. Section 2.3 provides insight into the different Speech bandwidth extension algorithms, while Section 2.4 outlines the research gap.

2.1 Production and Transmission of Speech

Understanding speech production and transmission is crucial for developing an effective Speech Bandwidth Extension(SBE) solution. Different sounds with unique characteristics are involved in speech production, and they can be grouped based on energy distribution across the frequency spectrum. Understanding these differences is essential for evaluating SBE algorithms and optimizing training to accurately predict speech sounds with high energy in the wide band (WB). The characteristics of speech signals transmitted through telephone networks are examined in terms of speech transmission. The goal is to minimize data trans-

mission while maintaining high performance at low costs. Initially, speech was transmitted at 8 kHz to meet the limitations of the NB (Narrowband) analog telephone standard. The transition to transmitting speech at 16 kHz significantly improved speech quality. However, the quality of decoded speech at the receiver’s end is also influenced by encoding at low bitrates. An effective SBE algorithm should perform well across different codecs used in NB speech transmission.

2.2 Impact of Resolution Limitations on Speech Bandwidth Extension and Restoration

The effectiveness of speech bandwidth extension and restoration depends heavily on the resolution of the input signal. Low-resolution signals inherently lack high-frequency components, which are critical for achieving high-quality and intelligible speech reconstruction. These limitations arise from fundamental principles of signal processing, including sampling constraints, aliasing, quantization, and band-limiting filters. Understanding these factors is essential for addressing the challenges in extending and restoring bandwidth, particularly in degraded or noisy conditions.

2.2.1 Sampling Rate and the Nyquist-Shannon Theorem

The Nyquist-Shannon Sampling Theorem states that to accurately reconstruct a signal, it must be sampled at a rate at least twice its highest frequency ($f_s \geq 2f_{max}$). This is known as the Nyquist rate. When the sampling rate falls below this threshold, the signal’s higher frequency components cannot be properly captured. In low-resolution signals, the sampling rate is often limited, which leads to the omission of frequencies above the Nyquist frequency. Consequently, these high-frequency components are effectively removed or misrepresented, degrading the signal’s fidelity and detail.

2.2.2 Aliasing and Its Effects

Aliasing is a phenomenon that occurs when high-frequency components of a signal are sampled at an insufficient rate, causing them to “fold” back into the lower frequency range. This happens because the sampling process cannot distinguish between the true frequency and its aliased counterpart. In low-resolution signals, where the sampling rate is low, aliasing becomes a significant issue. The result is a distorted representation of the signal, as high-frequency details are

either misrepresented as lower frequencies or lost entirely. Aliasing can severely compromise the integrity of the signal, making it challenging to reconstruct the original high-frequency content.

2.2.3 Quantization Resolution and High-Frequency Sensitivity

Quantization refers to the process of mapping a signal’s amplitude to discrete levels in digital systems. The resolution of this quantization is determined by the number of bits used to represent each sample. In low-resolution signals, fewer bits are available, which leads to larger rounding or truncation errors. High-frequency components, which often have smaller amplitudes and subtler variations, are more susceptible to these errors. As a result, the precision required to capture high-frequency details is lost, further degrading the signal’s representation.

2.2.4 Band-Limiting Filters in Signal Acquisition

Before sampling, signals are typically passed through a low-pass filter to limit their bandwidth. This process, known as band-limiting, ensures that no frequencies above the Nyquist frequency are present in the signal, thus preventing aliasing. However, in low-resolution systems with restricted sampling rates, the cutoff frequency of the band-limiting filter is lowered further to accommodate the limited resolution. This means that more high-frequency components are deliberately removed even before sampling occurs, leading to a signal that inherently lacks high-frequency information.

2.2.5 Granularity and Information Loss

High-frequency components require fine-grained temporal detail to be accurately represented. The granularity of a sampled signal is determined by how frequently the signal is sampled over time. In low-resolution systems, the sampling interval is too coarse to capture the rapid oscillations of high-frequency components. This insufficient granularity leads to the loss of detailed information, resulting in a signal that is smooth and oversimplified compared to the original. The inability to capture fine temporal details is a fundamental limitation of low-resolution signals.

In practical systems, low-resolution signals are often used in scenarios where computational or storage resources are constrained, such as in low-power devices or real-time systems. However, this trade-off comes at the expense of losing high-

frequency components, which can affect the clarity and quality of the signal. A useful analogy is capturing a rapidly oscillating sine wave with a coarse sampling rate. With a low resolution, only a few points are recorded along the wave, making it impossible to reconstruct its rapid oscillations. On the other hand, a higher resolution captures more points, preserving the wave’s fine details.

2.3 Speech Bandwidth Extension Algorithms

Speech bandwidth extension (SBE) tackles the challenge of expanding the bandwidth of signals that have been limited due to factors like imperfect transduction, compression algorithms, or noise interference[15]. Historical limitations in telecommunication design, influenced by technological and economic factors, often led to the adoption of band-limited versions of wide-band signals, despite advancements in microphone and digital signal processing technologies[16]. Early telecommunication systems, constrained by modem capabilities and storage limitations, prioritized restricting the bandwidth of speech signals while ensuring clarity[17]. SBE methods aim to estimate missing frequency components, which are classified into high-frequency or low-frequency categories based on the targeted frequency range[18]. Additionally, methods are categorized as blind or non-blind depending on whether prior information about the desired output is available[19]. For example, early 20th-century non-ideal transducers lacked prior information, while telecommunications systems could utilize filtered data to inform the estimation of missing spectral components. In modern digital signal processing, BWE techniques extensively leverage information from time, frequency, and joint time-frequency domains[20]. The advancement of Deep Learning has significantly contributed to addressing blind high-frequency BWE challenges[5]. Experimental setups typically involve down-sampling wide-band signals using anti-aliasing Low Pass Filters (LPF) followed by upsampling [16]. The resulting narrow-band signals are then used as inputs for BWE algorithms, aiming to reconstruct wide-band signals with enriched frequency content. Two primary BWE approaches are time domain and frequency domain methods [19]. Time domain BWE entails upsampling a downsampled wide-band signal to estimate missing high-frequency components. Conversely, frequency domain BWE involves estimating missing frequency bins and phase information within the Discrete Fourier Transform (DFT) domain[16].

2.3.1 Spectral Bands Mirroring

Spectral Bands Mirroring (SBM) was developed based on the notion that the spectral envelope of a signal contains the majority of its spectral information. This method exploits the strong correlation between the lower and upper halves of a signal’s Discrete Fourier Transform (DFT) magnitudes. By leveraging this correlation, SBR aims to replicate high-resolution signals through the translation of the lower half spectrum and normalization of the translated high-frequency spectrum with the spectral envelope. The frequency analysis in SBM relies on complex exponential-modulated quadrature mirror filter (QMF) banks. The SBM algorithm typically follows these steps:

1. Extraction of the spectral envelope from the input signal.
2. Filtering the signal with a Low Pass Filter (LPF) before transmission.
3. Transmission of both the filtered signal and the spectral envelope.
4. Transposing or mirroring the spectrum to obtain high-frequency spectral content.
5. Normalizing the spectrum with respect to the spectral envelope.

Initially designed for signal compression, SBM has been notably implemented in MPEG-1 Audio Layer 3 Pro (MP3Pro), enabling significant compression ratios. While SBM is a non-blind Bandwidth Extension algorithm, which falls beyond the scope of this thesis, a simplified version using Short-Time Fourier Transform (STFT) as a replacement for QMF filter banks is utilized for comparison purposes[21].

2.3.2 Approaches Utilizing Machine Learning

Machine learning approaches for high-frequency speech Bandwidth Extension (BWE) have evolved significantly since the pioneering work of Cheng et al. in 1994[22]. They observed that speech, characterized by phonemes or broad phonetic classes, exhibits distinctive patterns across the narrow-band or higher frequency spectrum. To avoid the complexities of classifying frames into phoneme classes, they assumed that each narrow-band sequence is produced by a sum of M random sources. Gaussian noise autoregressive modeling was employed for the sources, and the Expectation-Maximization algorithm was used for parameter estimation. Thereafter, in the same year, Yoshida and Abe[23]proposed a non-blind BWE algorithm based on codebooks, exploiting the high correlation

between narrow-band and wide-band spectral envelopes. This approach involved forming pairs of spectral envelopes to serve as queries and keys for the codebook. Each narrow-band frame’s spectral envelope was compared to those in the codebook, with the best match providing the index for the corresponding wide-band envelope. Building on Yoshida and Abe’s work, Epps and Holmes in 1999[24] categorized codebook-based BWE approaches into three categories: codebook mapping, codebook mapping with interpolations, and codebook mapping with split codebooks. They introduced the concept of separating codebooks into voiced and unvoiced counterparts. Innovative transformations, such as the Mel-frequency Cepstrum Coefficients (MFCC), have been proposed to enhance spectral representation[25]. Niklas and Enbom[26] replaced Linear Predicted power spectrum envelopes with MFCCs for BWE applications in 1999. Optimization techniques were explored between 2000 and 2010, focusing on spectral envelope estimation, codebooks, and vector quantization[27, 28, 29, 30, 31]. Iser and Schmidt[32] compared codebooks with Multilayer Perceptrons (MLPs) trained on Linear Predictive Coding (LPC) parameters, highlighting the importance of selecting an appropriate loss function for BWE tasks. Kontio et al. in 2007[33] introduced a partially recurrent fully connected MLP for BWE, incorporating spectral density energy distribution features into their model. They were the first to utilize reinforcement learning algorithms in BWE training. Li et al. in 2015[34] proposed the first frequency-based neural network architecture for BWE, utilizing Mel-transformed log-spectrograms as input. Liu and Bao[35] introduced Echo State Networks (ESN) as an alternative to discretized state-space models, demonstrating the effectiveness of continuous state-space modeling for BWE. Bachav et al. in 2017[36] adopted the constant-Q transform in place of the traditional Short-Time Fourier Transform (STFT), leveraging Gaussian Mixture Models to estimate log-scaled constant-Q transform coefficients.

2.3.3 Approaches Utilizing Deep Learning

In 2016, Deep Neural Networks (DNNs) made significant advancements in speech generation tasks, exemplified by WaveNet[41] and SampleRNN[42], which focused on raw audio modeling and frame-based solutions, respectively. These developments paralleled progress in computer vision, where super-resolution techniques addressed similar challenges to bandwidth extension (BWE) in speech processing. Dong et al.[43] proposed a two-layer CNN for end-to-end mapping between narrow-band and super-resolution images. Expanding on this, Kuleshov et al.[37] introduced the first end-to-end time-domain high-frequency BWE DNN, utilizing

an autoencoder with residual connections and a cubic-splined version of an aliased narrow-band signal. They introduced the Subpixel layer for upsampling instead of transposed convolution. Wang et al.[44] proposed using parallel WaveNets to model conditional probabilities and log-mel spectrograms for low-resolution audio.

Lim et al.[45] highlighted similarities between BWE and image super-resolution, proposing a DNN architecture with two parallel AudioUNets trained in frequency and time domains, integrating a spectral fusion layer. The spectral fusion layer equation is given as [46]:

$$M = w \odot |F(\hat{z})| + (1 - w) \odot \hat{m} \quad (2.1)$$

$$\hat{y} = F^{-1}(M \exp(j \cdot \text{angle}(F(\hat{z})))) \quad (2.2)$$

where F is the Fourier transform, $\hat{z}$ is the output of the time-domain section of TFNet, $\hat{m}$ is the frequency-domain section of TFNet, $\odot$ denotes element-wise multiplication, F^{-1} is the inverse Fourier transform, and w is a trainable parameter.

Building on these insights, Birnbaum et al.[38] advocated for using Recurrent Neural Networks (RNNs) to capture long-term dependencies, introducing batch normalization and Feature-Wise Linear Modulation (FiLM) for model stability and flexibility. Wang and Wang [47] proposed a linear combination of frequency and time loss components for enhanced performance using an Autoencoder CNN (AECNN) architecture.

Subsequent developments focused on lightweight CNN-based architectures to address computational resource constraints. Johnson and Parberry [48] proposed a stacked time-domain dilated CNN, while Dong et al. [49] incorporated attention mechanisms and non-local convolutions. Binqiang et al. [50] introduced a pyramid wavelet CNN (PWCNN) for DWT coefficient reconstruction, and Rakotonirima [39] proposed Attention FiLM (AFiLM) using Transformer blocks.

In 2020, Lagrange and Gontier [51] proposed a fully convolutional dilated neural network architecture for BWE, utilizing two non-dilated and L-2 dilated convolutional blocks. The architecture maintained consistent input, output, and feature map dimensions across the network. It employed dilation only on the frequency dimension, ensuring simplicity and interpretability, albeit with trade-offs in time-frequency resolution.

Time-domain solutions for BWE have proven the most successful, implicitly estimating both magnitude and phase. However, they require more trainable

parameters compared to frequency domain solutions. While efficient, frequency domain methods struggle with artifact reduction due to phase reconstruction errors. Table 2.1 presents a summary of a few recent works of deep learning-based speech bandwidth extension models concerning the domain, features, models, Contributions, and their limitations.

2.4 Research Gap

Despite the growing body of research on speech bandwidth extension, some issues such as handling noisy conditions, robustness to noisy variations, computational complexity, etc. have not been addressed to date. Based on the literature review on speech bandwidth extension (SBE) algorithms, several research gaps and unanswered questions emerge, pointing towards potential areas for further investigation. A significant research gap lies in the effective handling of noisy conditions during SBE for speech restoration. Although many algorithms show promising results under clean conditions, their performance often deteriorates in the presence of noise. Addressing this gap requires exploring robust SBE algorithms capable of accurately predicting high-frequency components even in noisy environments. Furthermore, the impact of different types and levels of noise on SBE performance remains unclear, raising questions about the robustness of existing algorithms in various noise scenarios.

Another research gap pertains to the computational complexity of SBE algorithms, especially in real-time applications. Many deep learning-based approaches exhibit high computational demands, limiting their practical deployment in resource-constrained environments such as mobile devices or embedded systems. Investigating lightweight architectures and optimization techniques tailored specifically for SBE could bridge this gap, enabling efficient and real-time implementation of bandwidth extension solutions.

In addition, the literature review highlights the need for comprehensive evaluation metrics that accurately reflect the perceptual quality of extended speech. Although traditional objective metrics such as SegSNR and LSD scores provide valuable insight into signal fidelity, they may not fully capture the subjective listening experience. Developing perceptually meaningful evaluation metrics tailored to the characteristics of extended speech could provide a more holistic assessment of the performance of SBE algorithms.

Table 2.1: Brief literature summary on a few recent works of deep learning-based speech bandwidth extension models concerning the domain, features, models, Contributions, and their limitations.

Article	Feature	Approach	Contribution	Limitation
Li et al., 2015[34]	STFT	Deep Neural Network(DNN)	Recover high band spectra in clean conditions, perform better in known phases, shows improved SegSNR and LSD scores.	Incorrect phase recovery, spectrum discontinuity, unable to recover high band in noisy conditions.
Kuleshov et al., 2017[37]	Waveform	Deep Convolutional Neural Network(CNN) with residual connection	Directly model natural raw signal, better scalability and computational resources.	Unable to handle large and complex dataset, unable to perform in noisy condition or distortion.
S. Birnbaum et al., 2019[38]	Waveform	Temporal Feature-Wise Linear Modulation(TFiLM)	Optimized Feature-Wise Linear Modulation(FiLM) for the use of one-dimensional convolutional layers.	Unable to capture long-term dependencies, unable to perform well in robust conditions.
Nathanael et al., 2021[39]	Waveform	Attention-based Feature-Wise Linear Modulation(AFiLM)	Efficiently incorporate long-range information, outperforms [37, 38] models on the multi-speaker task.	Arise computational complexity for more number of parameters, not well suited in noisy conditions.
Mandel et al., 2023[40]	STFT	AERO(Unet using complex-valued spectrogram + time and frequency domain loss function)	Recovery contributes in frequency axis only, can handle phase information efficiently, Directly predict full frequency range for reconstruction.	Complex valued spectrogram introduce additional computational complexity, <128M parameter create potential limitation.

Chapter 3

Joint Speech Bandwidth Expansion and Restoration

This chapter¹ explores a joint approach, aiming to simultaneously expand bandwidth and restore speech quality using Deep Neural Networks (DNNs). With the primary objectives of recovering high-frequency spectra and reducing noise in speech signals, the research proposes a DNN-based approach to reconstruct clean full-band speech from noisy low-band signals. By addressing these challenges, the study seeks to advance speech processing technologies, benefiting speech communication, recognition, and synthesis. Motivated by the limitations of traditional methods, the research harnesses the potential of DNNs to innovate in the field of speech processing.

Section 3.1 provides insight into the motivation behind the proposed framework and reviews related studies. Section 3.2 outlines the proposed architecture, while Section 3.3 details the experimental setup and performance evaluation across various scenarios. Finally, Section 3.4 summarizes the chapter, highlighting the key findings and contributions of the research.

3.1 Motivation

In the realm of speech processing, the joint expansion of bandwidth and restoration of speech quality has garnered significant attention. The primary challenge lies in simultaneously augmenting the bandwidth of speech signals while improving their fidelity. Traditional methods have struggled to achieve this dual objective effectively. However, with the advent of Deep Neural Networks (DNNs), there's a newfound promise in addressing this challenge. DNNs exhibit the ca-

¹The works in this chapter have been published in [52]

capacity to comprehend intricate relationships between input and output signals, making them suitable candidates for tackling the complexities inherent in joint bandwidth expansion and speech enhancement tasks. Recent advancements have led to the development of algorithms that leverage DNNs for this purpose. These algorithms aim to not only increase the bandwidth of speech signals but also to diminish noise, all while preserving speech quality and intelligibility. By harnessing the capabilities of DNNs, researchers have proposed innovative approaches to jointly expand bandwidth and restore enhanced speech. The motivation behind this study stems from the recognition of the limitations of existing methods and the potential offered by DNN-based approaches. By exploring the possibilities afforded by DNNs, there's an opportunity to revolutionize speech processing techniques and overcome longstanding challenges in bandwidth expansion and speech restoration. The aim of this thesis is twofold-

- Firstly, it aims to tackle the challenge of bandwidth expansion, endeavoring to recover the high-frequency spectrum of speech signals.
- Secondly, it seeks to address the issue of noise reduction, aiming to reconstruct clean speech from noisy speech signals.

To achieve these objectives, the study proposes a DNN-based approach designed to reconstruct clean full-band speech spectra from their noisy low-band counterparts. Through this research, we aspire to contribute to the advancement of speech processing technologies, particularly in the domains of speech communication, speech recognition, and speech synthesis.

3.2 Proposed Architecture

The proposed system depicted in Figure 3.1 comprises two main stages: training and testing. In the training stage, a full-band audio signal is initially filtered to obtain a narrowband signal using a 20th-order Butterworth filter with a cutoff frequency of 3000 Hz. Subsequently, short-time Fourier transform (STFT) is applied to extract features from the audio signal. During training, the model takes clean narrowband audio as input and aims to predict the corresponding clean full-band audio signal. For extracting features, targeted wideband signal is segmented into frames using STFT. Equation 3.1 illustrates the STFT process, where each frame is represented as $Y(m,n)$, with m denoting the frame index, n representing the frequency index, Δ indicating the window shift, N being the window length, and $h(k)$ representing the Hamming window function. From the STFT, log-spectral power magnitude and phase information are extracted as

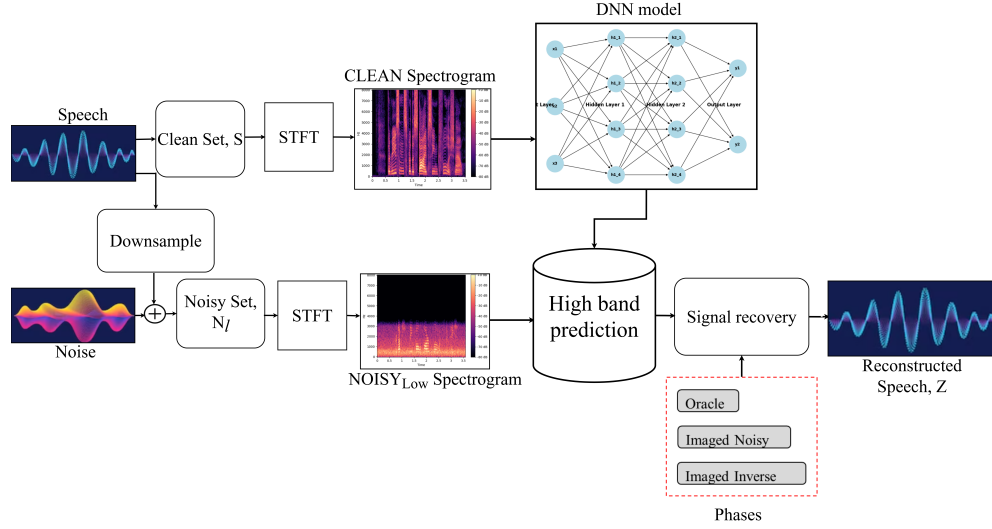


Figure 3.1: Overall block diagram of proposed DNN-based joint speech bandwidth expansion and speech restoration.

features. During testing, a noisy narrowband signal is inputted into the model to recover the clean full-band signal. The input narrowband signal, x , is generated by filtering and downsampling the wideband signal y . Features of the narrowband signal, denoted as X_{mag} and X , represent the log-spectral magnitude and phase, respectively.

$$Y(m, n) = \sum_{k=0}^{N-1} y(m * \Delta + k) h(k) e^{-2\pi j k n / N} \quad (3.1)$$

3.2.1 Model Description

The DNN architecture, depicted in Figure 3.1, consists of one input layer, one output layer, and three hidden layers, each containing 256 nodes. The input to the DNN comprises the log spectrum of the narrowband signal, while the output is the high-frequency log spectrum of the wideband signal. To ensure the effectiveness of the DNN, the input and output dimensions are normalized across all training samples to have zero mean and unit variance. This normalization process is also applied during the bandwidth expansion stage, where the input feature vectors are normalized, and the output undergoes reverse normalization. The DNN operates with a window size of 256 and a window shift of 128 for the narrowband signal. The mean squared logarithmic error (MSLE) is utilized as the training loss function, with a base learning rate set to 10^{-5} . The 'Adam' optimization method is employed with batches of 64 examples and a learning rate of 0.001. Training halts when the MSLE value remains unchanged for five consecutive epochs, indicating convergence to an optimal condition. The MSLE

calculation, as described in Equation 3.2, quantifies the difference between the log-transformed actual signal (y) and the predicted signal ($\hat{y}$) across all input signals (n). This error measure guides the training process toward minimizing the discrepancy between the predicted and actual signals.

$$E_i = \sum_{i=1}^n (\log(y_i + 1) - \log(\hat{y}_i + 1))^2 \quad (3.2)$$

Here, n represents the number of input signals, y_i denotes a data point of the original signal, and $\hat{y}_i$ is the data point value of the predicted signal.

3.2.2 Reconstruction of Signal

Signal reconstruction focuses on restoring the wideband signal from the predicted high-frequency spectra while addressing the challenge of phase information. Various techniques have been developed to estimate the phase for high-frequency components predicted by the model. The high-frequency spectra are derived from the output of the DNN and compensate for energy losses caused by the use of only half the points from the wideband signal to construct the narrowband spectrum [53].

To reconstruct the phase information, three conditions are considered:

1. Oracle Phase:

$$\angle Y = \angle \hat{Z} \quad (3.3)$$

Here, $\angle Y$ represents the clean full-band phase, which is directly used as the reconstructed phase ($\angle \hat{Z}$).

2. Imaged Phase:

$$\angle \hat{Z} = -\text{flip}(\angle X_{\text{low}}) \quad (3.4)$$

The phase is obtained by flipping and negating the noisy narrowband phase ($\angle X_{\text{low}}$).

3. Noisy Phase:

$$\angle X_{\text{noisy}} = \angle \hat{Z} \quad (3.5)$$

The reconstructed phase ($\angle \hat{Z}$) is taken directly from the noisy full-band phase ($\angle X_{\text{noisy}}$).

In the testing phase (illustrated in Figure 3.1), the input noisy narrowband signal is processed using windowing via the short-time Fourier transform (STFT) to estimate the high-band signal. After estimating the magnitude of the high-frequency signal, waveform reconstruction is performed by combining the estimated magnitude with the respective phase information under the mentioned

conditions.

The reconstructed signal is computed as:

$$\hat{Z} = \hat{Z}_{\text{mag}} \cdot \exp(j \cdot \angle \hat{Z}) \quad (3.6)$$

where $\hat{Z}_{\text{mag}}$ is the estimated magnitude, and $\angle \hat{Z}$ is the reconstructed phase.

Finally, the inverse short-time Fourier transform (ISTFT) is applied to $\hat{Z}$, followed by overlapping and adding with the same Hamming window to reconstruct the signal accurately. This approach ensures proper phase and magnitude integration for high-quality wideband signal reconstruction.

3.3 Experiments and Results

3.3.1 Dataset Description

This experiment has been conducted by using TIMIT Acoustic Phonetic Continuous Speech Corpus[54], sponsored by DARPA-ISTO, is a comprehensive collection of read speech designed for acquiring acoustic-phonetic knowledge and developing and evaluating automatic speech recognition systems. It was developed collaboratively by several institutions, including MIT, SRI, and TI, with speech recorded at TI, transcribed at MIT, and managed by NIST. The corpus consists of 6300 sentences, with each of the 630 speakers delivering 10 sentences, representing 8 major dialect regions of the United States. The corpus includes a diverse range of speech material, such as dialect sentences, phonetically-compact sentences, and phonetically-diverse sentences. The corpus is subdivided into training and testing data, with a core test set containing 24 speakers and a more extensive complete test set comprising 168 speakers. The corpus also includes associated files for each utterance, including speech waveform files, orthographic transcriptions, word transcriptions, and phonetic transcriptions. Online documentation provides detailed information on corpus design, collection, and transcription.

3.3.2 Experimental Setup

This study utilized 1000 samples sourced from various male and female speakers from the TIMIT corpus for training, alongside 50 samples for testing. All speech samples were formatted as 16-bit, 16kHz waveform files. Separate sets of speakers were used for training and testing, with 20% of the training data reserved for validation. All speech samples were standardized to a uniform sampling rate of 16kHz. By training on clean signals, the model effectively learns the fundamen-

tal characteristics of speech, allowing it to generalize well and reconstruct high-quality wideband signals, even in noisy conditions. During testing, this training strategy enables the model to leverage these learned characteristics to enhance speech from noisy inputs, as reflected in improved numerical values of evaluation metrics such as SNR, PESQ, and STOI. To evaluate the model’s robustness, testing was conducted using both clean and noisy signals. This approach allowed for a comprehensive assessment of the model’s efficacy in reconstructing wideband signals under realistic conditions where noise significantly impacts speech signals. The experiments were performed on the TIMIT corpus, with audio sampled at 16kHz and 16-bit resolution. Short Time Fourier Transform (STFT) was applied with a window size of 256 samples and a shift length of 128 samples for both wideband and narrowband signals. For training, the Mean Squared Logarithmic Error (MSLE) was used as the loss function, with a base learning rate set to 10^{-5} . The Adam optimization algorithm was employed with batch sizes of 64 examples, and training was terminated when the loss value remained constant for five consecutive epochs. Additionally, Minimum Mean Squared Error (MMSE) training was conducted to further optimize the model. This phase also used a base learning rate of 10^{-5} , with the Adam optimizer dynamically adjusting the learning rate based on the reduction in mean squared error. MMSE training adopted a mini-batch strategy with a batch size of 32 utterances and halted once the mean squared error reached a minimum threshold of 0.001. By leveraging clean signals for training, the model was designed to adaptively learn the fundamental characteristics of speech, making it more capable of recovering signals from noise during testing. This approach highlights the importance of model adaptability for achieving robust signal reconstruction in both clean and noisy conditions.

3.3.3 Results

This study evaluated the performance using various objective measures, comparing the enhanced speech signal with the clean reference from the TIMIT test set. To simulate real-world scenarios, two types of noise, namely speech-shaped noise and car noise, were added to the clean signal at different signal-to-noise ratio (SNR) levels. Objective metrics such as perceptual evaluation of speech quality (PESQ), short-time objective intelligibility (STOI), and spectrogram orthogonal polynomial measure (SOPM) were computed to assess the quality and intelligibility of the enhanced speech signal.

- **Perceptual Evaluation of Speech Quality (PESQ):** PESQ is a widely used objective metric for evaluating the quality of speech signals. It mea-

asures the similarity between the original and degraded speech signals by simulating human perception. PESQ scores range from -0.5 to 4.5, where higher scores indicate better speech quality. A score close to 4.5 indicates excellent quality, while a negative score suggests severe degradation[55].

- **Spectrogram Orthogonal Polynomial Measure (SOPM):** SOPM is a metric used to assess the quality of enhanced speech signals. It quantifies the similarity between the spectrograms of the original and enhanced signals, focusing on preserving spectral characteristics. SOPM scores typically range from 0 to 1, where a score closer to 1 indicates better preservation of spectral features. A score of 0 implies significant distortion or degradation[56].
- **Short-Time Objective Intelligibility (STOI):** STOI measures the intelligibility of speech signals by analyzing their temporal characteristics. It evaluates how well speech can be understood in noisy environments by comparing the similarity of short-time segments between the original and degraded signals. STOI scores range from 0 to 1, with 1 indicating perfect intelligibility and 0 representing complete unintelligibility. Scores closer to 1 suggest better speech intelligibility in noisy conditions[57].

3.3.3.1 Objective scores in Clean scenario

Table 3.1 displays the outcomes of an ablation study conducted on reconstructed signals from the TIMIT corpus under clean conditions. This study contrasts the original signal with three distinct reconstructed signals featuring different phases (Oracle, Imaged, and Noisy). The results shed light on the quality and intelligibility of the reconstructed speech signals across various phases. Notably, the signal reconstructed using the oracle phase exhibits the highest PESQ score of 3.04, indicating commendable speech quality. Following closely are the noisy phase at 2.98 and the imaged phase at 2.96. In terms of STOI, the oracle phase achieves the highest intelligibility score of 0.82, trailed by the noisy phase at 0.78 and the imaged phase at 0.76. Additionally, SOPM scores demonstrate that the oracle phase effectively preserves spectral characteristics, scoring 0.86, while the imaged phase follows at 0.79 and the noisy phase at 0.76. Overall, although all phases exhibit relatively good performance across the objective measures, the oracle phase consistently outperforms others in terms of both speech quality and spectral preservation.

Table 3.1: Effects of different phases in Clean Scenario.

Objective Measures	Noisy Low Signal	Reconstructed Signal		
		Oracle phase	Imaged phase	Noisy phase
PESQ	1.62	3.04	2.96	2.98
STOI	0.77	0.82	0.76	0.78
SOPM	0.75	0.86	0.79	0.76

Table 3.2: Performance Evaluation of Full Band Noise (SSN)

Objective Measures	10 dB				5 dB			
	Unprocessed	Enhanced			Unprocessed	Enhanced		
		Oracle	Imaged Inverse	Imaged Noisy		Oracle	Imaged	Noisy
PESQ	2.3	2.6	2.32	2.32	1.97	2.31	1.97	1.97
STOI	0.86	0.9	0.86	0.87	0.77	0.84	0.79	0.78
SOPM	0.92	0.96	0.93	0.92	0.85	0.9	0.86	0.86

3.3.3.2 Objective scores in Noisy scenario

Table 3.2 and Table 3.3 present the results of an evaluation of audio signal processing techniques in two different noisy environments: full-band noise (SSN) and low-band noise (Car). Each table contains three objective measures: PESQ, STOI, and SOPM. The results are provided for both the noisy signal and the enhanced signal under different conditions: using the oracle phase, imaged phase, inverse phase, and noisy phase. Comparing the performance of the noisy signal with the enhanced signal, it is evident that the enhanced signal generally outperforms the noisy signal across all objective measures. In Table 3.2, the full-band noisy environment (SSN), the enhanced signal shows improvement in PESQ, STOI, and SOPM values by 30%, 4%, and 4% respectively, compared to the noisy signal at 10 dB, and by 34%, 7%, and 5% respectively at 5 dB. Similarly, in Table 3.3, in the low-band noisy environment (Car), the enhanced signal demonstrates improved performance metrics compared to the noisy signal, particularly at lower SNR levels. Overall, these tables highlight the effectiveness of the proposed system in enhancing speech signals in noisy environments, with the enhanced signal exhibiting higher quality, intelligibility, and spectral preservation compared to the original noisy signal.

Table 3.3: Performance Evaluation of Low Band Noise (Car)

Objective Measures	10 dB				5 dB			
	Unprocessed	Enhanced			Unprocessed	Enhanced		
		Oracle	Imaged Inverse	Imaged Noisy		Oracle	Imaged Inverse	Imaged Noisy
PESQ	2.32	2.6	2.32	2.32	1.97	2.3	1.97	1.97
STOI	0.86	0.9	0.88	0.88	0.76	0.82	0.78	0.77
SOPM	0.91	0.95	0.93	0.923	0.85	0.9	0.86	0.86

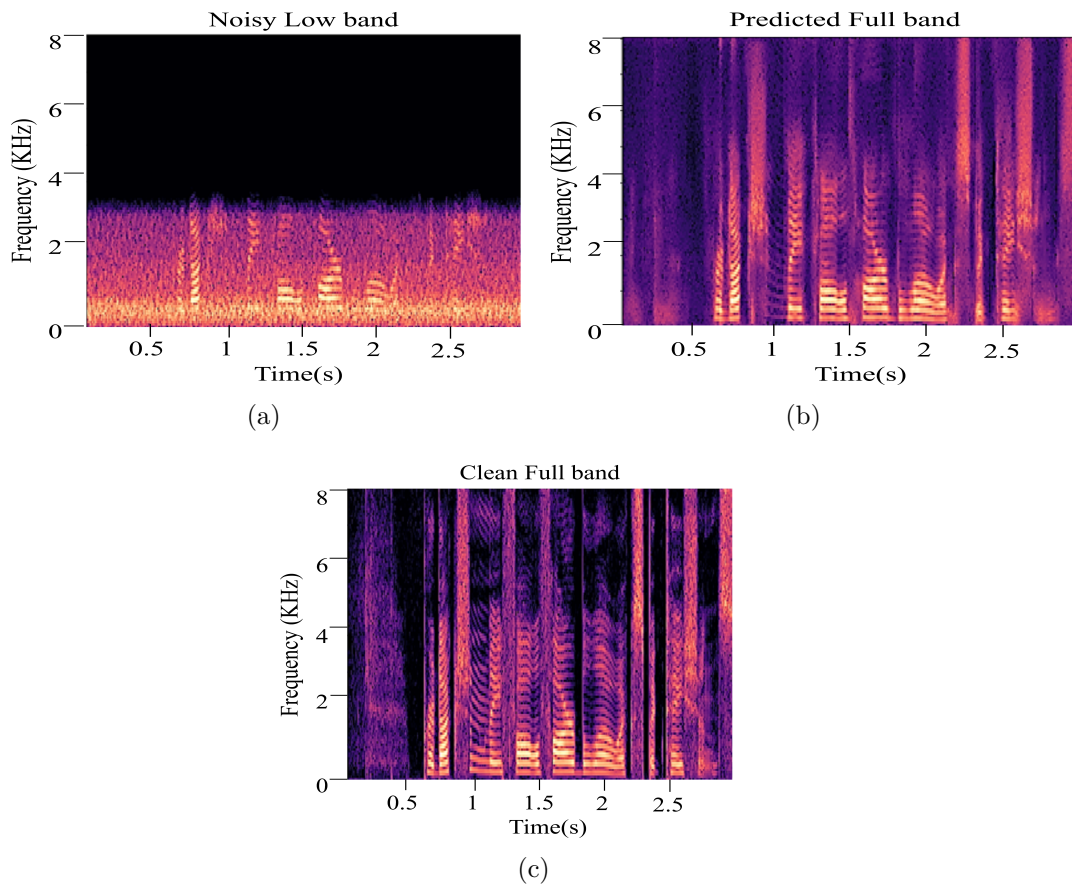


Figure 3.2: Spectrogram of one test utterance with SNR 10 dB, (a) Noisy low band signal spectrogram, (b) Predicted full band signal spectrogram, (c) Original full band signal spectrogram. .

3.3.3.3 Spectrogram analysis

Figure 4.5 provides a visual representation of a male utterance extracted from the TIMIT test set. It is evident that when directly predicting the high-frequency spectrum, certain high-frequency bands are obscured or washed out. However, the proposed approach is capable of reconstructing the full-band spectrogram without any discontinuity.

3.4 Summary

In this chapter, a novel deep neural network (DNN) based approach is presented for simultaneous bandwidth expansion and speech restoration. The primary objective is to enhance the quality of speech signals while increasing their bandwidth and reducing noise interference. The network incorporates three distinct phases, leveraging magnitude spectra to recover high-band components. Its design allows for the concurrent estimation of missing speech components and noise profiles

within degraded speech signals. Results indicate that the DNN-based approach excels in noisy environments, surpassing conventional methods. Through a comparative and evaluation of audio signal processing, the efficacy of the enhanced signal is demonstrated compared to the original noisy low-band signal. Notably, the reconstructed signal with the "oracle" phase exhibits superior performance, achieving a PESQ score of 3.04 and SOPM score of 0.86. While the "imaged" and "noisy" phases also yield comparable results, the oracle phase stands out. The improvement is more pronounced in full-band noisy environments, where the enhanced signal enhances PESQ, STOI, and SOPM values by 0.04, 0.03, and 0.01 respectively at 10 dB compared to 5 dB. Similarly, in low-band noisy environments, enhancements of 0.05, 0.02, and 0.03 are observed at 10 dB compared to 5 dB. These findings underscore the potential of the proposed method to effectively enhance speech signal quality in noisy settings.

Overall, while the reconstructed signal quality remains robust in clean scenarios, there is room for improvement in noisy conditions. Particularly, the performance of the oracle phase indicates promising prospects for further refinement. Further investigation into the effects of different signal-to-noise ratios (SNRs) across various noisy conditions is warranted to validate the method's robustness comprehensively.

Chapter 4

RDNet: Robust extension-plus-enhancement of speech utilizing Dual-former Network

This chapter addresses the challenge of mitigating mismatch issues resulting from channel effects in noisy environments. In Section 4.1, the chapter discusses the motivation behind the study and reviews related research. Section 4.2 presents a novel approach that integrates a bandwidth extension module with a speech enhancement module to address scenarios where high-frequency information is lost due to channel effects in noisy conditions. The experimental setup and results are detailed in Section 4.3, while Section 4.4 provides a summary of the chapter’s findings.

4.1 Motivation

In real-world scenarios, speech signals often suffer from distortion caused by background noise and the loss of high-frequency information due to transmission channels. As wider bandwidths in speech signals lead to better perceptual quality and intelligibility, recovering missing frequency information becomes crucial for tasks like speech enhancement, especially when facing the mismatch problem due to channel effects.

Bandwidth extension techniques are commonly employed to recover high-frequency information from narrowband signals, benefiting applications such as hearing aids design [9, 58] , speech recognition [34, 59, 60], and speaker verifi-

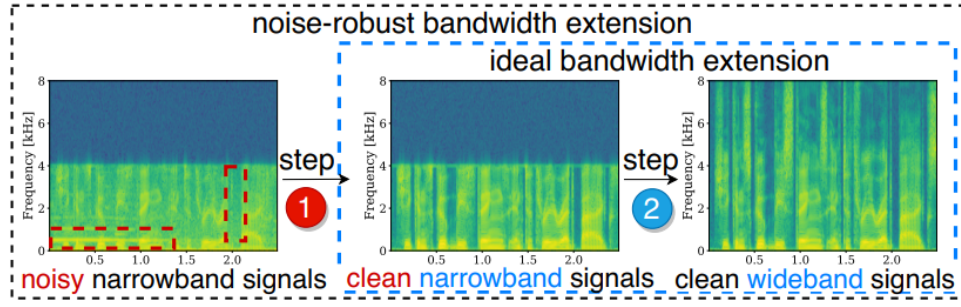


Figure 4.1: The process involves two key steps: First, the noisy narrowband signal undergoes enhancement to eliminate noise (Step 1). Second, the improved narrowband signal is extended in bandwidth to produce the clean fullband signal (Step 2) [1].

cation [61, 62]. However, existing bandwidth extension methods, including deep neural networks (DNN) [34, 63], fully convolutional networks [37, 64], generative adversarial networks (GAN) [65], and WaveNet [66], primarily operate under ideal conditions with clean narrowband signals as inputs, termed as ideal bandwidth extension. However, in real-world scenarios where signals are distorted and partially missing due to channel effects, the conventional bandwidth extension approach alone cannot address this condition effectively. A common approach to recover high band spectra involves performing speech enhancement on the noisy narrowband signal first (Step 1) and then conducting ideal bandwidth extension (Step 2) shown in Figure 4.1. For instance, P. J. Moreno et. al, has employed iterative Vector Taylor Series (VTS) approximation algorithms [67] for feature enhancement, followed by Gaussian mixture models or maximum a posteriori models to reconstruct wideband signals [68, 69]. Recently, there has been a trend towards using a single neural network for combining speech enhancement and bandwidth extension (UEE) using deep learning techniques. This method has become well-liked because it is both efficient and effective. The UEE approach typically involves a two-stage training scheme where speech enhancement and bandwidth extension modules are trained separately and then fine-tuned jointly with a single mean square error (MSE) loss between the clean wideband ground-truth and the enhanced-plus-extended output [70].

This chapter proposes an end-to-end frequency-domain framework realized through integrating speech enhancement and ideal bandwidth extension modules within a Robust extension-plus enhancement of speech utilizing Dual-former Network (RDNet). Our method aims to recover full-band speech from noisy low-band signals directly. Specifically, the noisy narrowband signal is initially encoded into short-time Fourier transform (STFT) features. The speech enhancement module

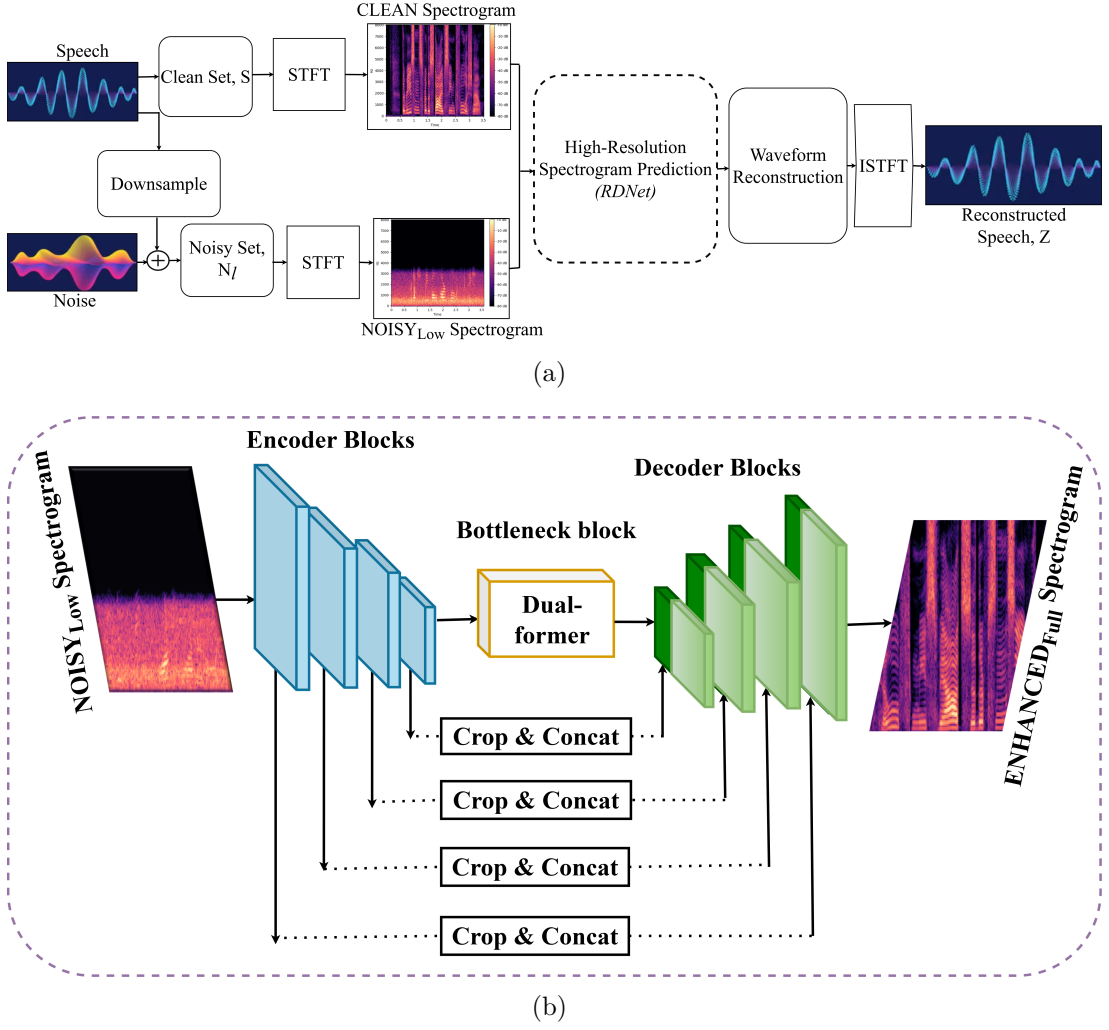


Figure 4.2: An illustration of the proposed network architecture(RDNet). The processing chain depicted in (a) illustrates the comprehensive prediction core, utilizing the Short-Time Fourier Transform (STFT) signal transformation, high-resolution spectrogram prediction, and the Inverse STFT process. Meanwhile, the modular components illustrated in (b) represent individual cores, each equipped with learned feature transformations based on 2D-Conv encoder-decoder layers, and a dual-former layer as bottleneck.

then utilizes these STFT features to estimate spectrograms and obtain enhanced full-band features for subsequent bandwidth extension. Finally, speech decoders are trained to reconstruct the enhanced full-band and enhanced-plus-extended features into frequency-domain signals using inverse STFT (iSTFT).

4.2 Proposed Architecture

This section introduces a frequency-domain bandwidth extension module with enhancement capabilities, as depicted in Figure 4.2. We begin by exploring a

network designed for both enhancement and extension in the frequency domain. This network comprises a 2-D convolutional encoder and a 2-D de-convolutional decoder, as depicted in Figure 4.2(b). Such a convolutional encoder-decoder structure is commonly utilized in enhancement and separation tasks [71]. The enhancement and extension processes are carried out sequentially using two deep learning networks: the encoder-decoder serves as the backbone, while the dual-former acts as the bottleneck.

4.2.1 Model Description

The system operates by first applying the Short-Time Fourier Transform (STFT) to the clean audio signal, generating a spectrogram that encapsulates its spectral content. Simultaneously, a noisy version of the signal is created through down-sampling and noise introduction, and its spectrogram is sent to the model alongside the clean spectrogram. Following model processing, waveform reconstruction merges spectral information from both signals, preserving phase data through the Inverse Short-Time Fourier Transform (ISTFT) to yield a reconstructed signal. The resulting output is an enhanced version of the original, enriched by the model’s restoration process, enhancing signal quality and fidelity.

The model architecture features four encoder layers utilizing convolutional operations for precise feature extraction, while a pivotal bottleneck layer (Dual-former) comprising a transformer and conformer reshapes encoder outputs to prepare for subsequent processing. Decoding involves up-sampling, skip connections, convolutional transpose, and normalization techniques to faithfully reconstruct the input shape. Skip connections facilitate seamless information flow between encoder and decoder layers, mitigating potential information loss. The encoder-decoder architecture, favored for tasks like speech enhancement and bandwidth extension, excels in capturing and reconstructing intricate temporal and spectral patterns within audio signals. Through its sequence-to-sequence modeling approach, the encoder adeptly extracts pertinent features from the input, while the decoder utilizes these representations to faithfully reconstruct the desired output. This hierarchical representation learning enables the model to effectively abstract both short-term dynamics and long-term contextual information, ensuring robustness and accuracy across diverse input conditions, making it a popular choice for such applications, often outperforming shallower network architectures like DNNs, CNNs, or RNNs. Its inherent flexibility allows it to accommodate varying input lengths, multi-channel audio, and additional contextual cues, making it a versatile and potent solution for enhancing speech quality and extending

bandwidth in real-world applications.

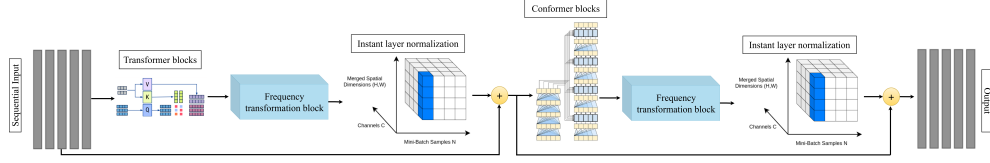


Figure 4.3: Architecture of Dual-former block consisting of a transformer, a conformer, frequency transformation and instant layer normalization layers.

4.2.2 Dual-former

In Figure 4.3 a dual-former architecture is shown, which is a critical component for the proposed approach. The functionality and structure of this architecture along with the relevant equations has been broke down below.

1. **Transformer block:** The output from the last encoder layer is reshaped into a suitable format for processing and then fed into a transformer layer shown in Figure ??, which consists of multi-head attention mechanisms with four heads and feed-forward neural networks [39]. This layer captures complex dependencies within the input sequence. The output of the transformer layer undergoes additional processing, including normalization and concatenation with the original input. Finally, the concatenated output is reshaped and passed through a normalized layer to produce the final transformer path output. The equations for the transformer encoder can be summarized as follows:

- **Multi-Head Self-Attention:**

$$\text{Attention}(Q, K, V) = \text{softmax} \left(\frac{QK^T}{\sqrt{d_k}} \right) V \quad (4.1)$$

Here: Q (Query): Represents the current element's "query" about other elements in the sequence. It is a transformation of the input sequence. K (Key): Acts as the "key" for each position, used by the query (Q) to determine relevance. It is also a transformation of the input sequence. V (Value): Represents the actual content or data corresponding to each position in the sequence. It is another transformation of the input sequence. d_k : Refers to the dimensionality of the keys (and queries). It is used to scale the dot product QK^T , ensuring numerical stability during computation.

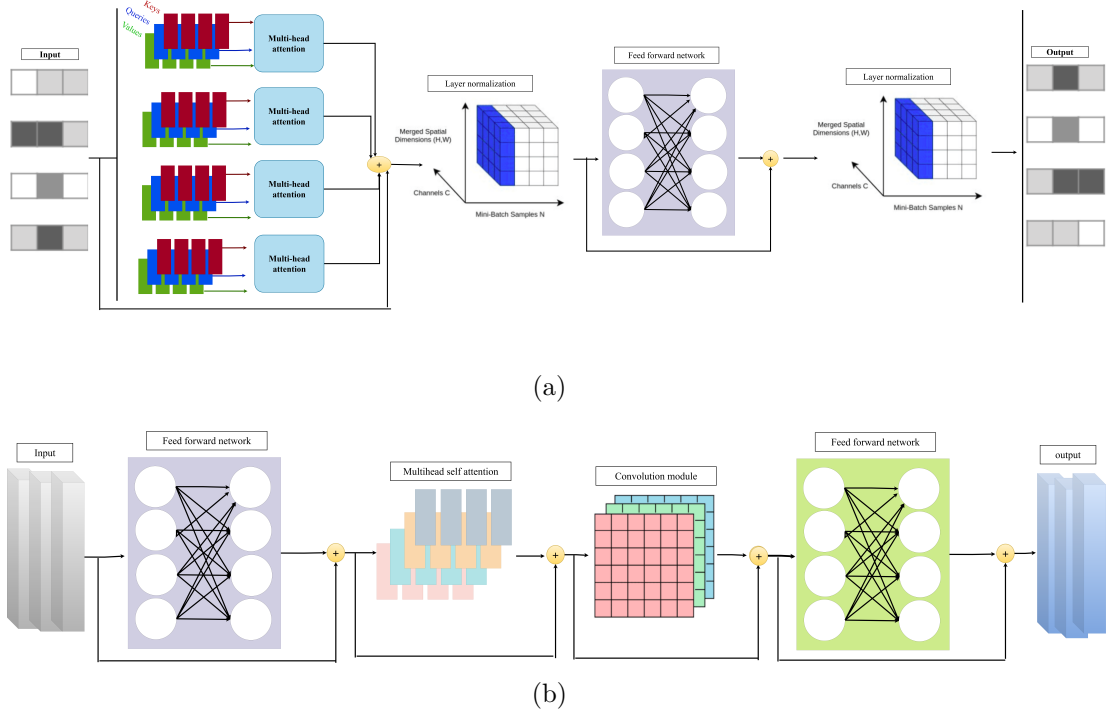


Figure 4.4: An illustration of (a) transformer and (b) conformer layer used in Dual-former block.

This mechanism computes the relevance of each sequence position to the current position using the dot product QK^T . The attention weights are normalized using the softmax function, which assigns a probability distribution across the sequence. These weights are then used to compute a weighted sum of the values (V) to generate the attention output.

- **Feed-Forward Networks:**

$$\text{FFN}(x) = \text{ReLU}(xW_1 + b_1)W_2 + b_2 \quad (4.2)$$

Here: x : The input to the feed-forward layer, typically the output from the attention mechanism. W_1, W_2 : Learnable weight matrices. W_1 transforms the input into a higher-dimensional space, while W_2 projects it back into the original or desired dimensionality. b_1, b_2 : Bias terms added during each linear transformation to increase the model's flexibility. ReLU: A non-linear activation function, introducing non-linearity to enhance the model's representational capacity.

The transformer captures long-range dependencies and relationships between elements in the sequence through self-attention. The feed-forward network further refines the output by applying non-linear transformations

and dimensionality changes. Together, these components enable the model to learn and represent complex dependencies efficiently.

2. **Conformer block:** Drawing inspiration from the Convolution-augmented Transformer proposed by Gulati et al. [72] for speech recognition, a modified version of the transformer encoder, incorporating additional convolutional layers shown in Figure ?? . This architecture enhances the model’s capability to capture both local and global dependencies in the input sequence. The conformer block integrates multiple components, including multi-head attention with four heads, feedforward neural networks (FFN), and convolutional neural networks (CNN). Specifically, our conformer block architecture comprises two feedforward modules surrounding the multi-headed attention module and the convolution module. This structural design, reminiscent of the Macaron-Net architecture [73], forms the core framework. In contrast to the original Conformer proposed by Gulati et al. [72], which employed two Macaron-net style feed-forward layers with half-step residual connections, RDNet adapts and utilizes two vanilla-style feed-forward layers with residual connections. This arrangement, with feed-forward layers strategically positioned around the attention and convolution modules, yields significant enhancements, improving the block’s ability to capture intricate sequential relationships. Inspired by the Macaron-Net structure, our proposed modification introduces a novel approach to organizing feed-forward layers, thereby augmenting the overall effectiveness of the Conformer block in modeling complex dependencies within sequential data.
3. **Instant layer normalization:** a layer normalization technique, which helps stabilize the training process and improve model performance. The equations for instant normalization are as follows:

$$\begin{aligned}
 \text{mean} &= \text{mean}(x) \\
 \text{variance} &= \text{mean}((x - \text{mean})^2) \\
 \text{normalized_x} &= \frac{(x - \text{mean})}{\sqrt{\text{variance} + \epsilon}} \\
 \text{output} &= \text{normalized_x} \times \text{gamma} + \text{beta}
 \end{aligned} \tag{4.3}$$

The dual-former block implements the architecture by combining the transformer and conformer blocks. This architecture utilizes two parallel paths for feature extraction, enabling the model to capture both spectral and temporal information

effectively. Overall, the dual-former architecture provides a powerful framework for processing sequential data, such as audio signals, by leveraging the strengths of both transformer and convolutional operations.

4.3 Experiments and Results

4.3.1 Dataset Description

This research aims to assess the effectiveness of the proposed RDNet model through experiments conducted on the following datasets:

- **TIMIT database**¹ [54]: The TIMIT database consists of 6,300 speech utterances in American English, each accompanied by phonetic transcriptions, with durations ranging from 3 to 5 seconds. To create training dataset, a subset of 500 utterances from 50 different speakers from the TIMIT database is selected. These sentences were then modified by filtering out high-frequency components and introducing noise from four distinct sources, resulting in the creation of the Noisy Lowpass dataset. This dataset includes three Signal-to-Noise Ratios (SNRs): -5, 0, and 5 dB. Within the training dataset, 150 randomly selected utterances to form a validation test set is set aside. The environmental noise conditions encompass various sources such as airports and speech-shaped noise (SSN). Notably, airport is classified as low-band, while SSN belong to the high-band or mid-band category. For model evaluation, a secondary subset comprising 50 samples for testing is utilized. These utterances were mixed with two "seen" noise types (Airport, SSN) across three different SNR levels: -5, 0, 5 dB. All speech samples and noises were resampled at a rate of 16 kHz.
- **DNS Challenge dataset:**The DNS Challenge (INTER-SPEECH 2020) dataset, as referenced in [74], is a comprehensive collection designed for speech enhancement research. It encompasses a clean speech dataset comprising more than 500 hours of recordings from 2150 speakers. Additionally, the noise dataset contains over 180 hours of audio clips spanning 150 distinct classes. To maximize the dataset's utility, a dynamic mixing approach is employed during model training. Specifically, a training subset is created by randomly selecting 100 utterances and combining them with five randomly chosen noises from both free sound and audioset sound categories. During each training epoch, these clean speeches are mixed with

¹discussed in section 3.3.1

the selected noises, resulting in speech-noise mixtures with varying signal-to-noise ratios (SNRs) ranging from -5 dB to 5 dB. Notably, the mixing process is conducted dynamically, ensuring diverse combinations for robust model training. For evaluation purposes, the DNS Challenge provides a publicly available test dataset comprising synthetic clips categorized into two groups: those without reverberations and those with reverberations. Each category contains 150 noisy clips with SNR levels distributed between 0 dB and 20 dB. To assess model performance, a secondary subset consisting of 10 samples is utilized for testing. These samples are mixed with five random "seen" noise types across different SNR levels, including both seen (-5 dB, 0 dB, and 5 dB) and unseen (-2.5 dB, 2.5 dB, and 7.5 dB) levels. Furthermore, all speech samples and noises in the dataset are resampled at a rate of 16 kHz to ensure consistency across the dataset.

4.3.2 Experimental setup

The experimental setup commences by initializing the Adam optimizer with a learning rate set to 0.001. This optimizer configuration is then utilized to compile the proposed model, where the Mean Absolute Error (MAE) is chosen as the loss function. Custom Data Generator instances are employed to efficiently generate batches of training data during model training. The training process spans 30 epochs, with the number of steps per epoch determined by dividing the length of the input dataset by the batch size. Following the training phase, model evaluation is performed on the test data using the evaluate method. This involves computing loss values and metrics on the provided test data, with the resulting metrics presented for further analysis. For the testing and reconstruction phase, a frame reconstruction function is defined to reshape the model predictions and input test data to the appropriate dimensions. The model's prediction method is then applied to generate predictions on the test data. Subsequently, the resulting magnitude predictions undergo reconstruction using the defined function. Additionally, reconstruction of the input and target phases is executed to facilitate subsequent computations. Depending on the specified phase (e.g., 'oracle'), full phase information is reconstructed accordingly. The predicted magnitude matrix is multiplied by the exponential of the full phase information, followed by the application of the inverse short-time Fourier transform (iSTFT) to reconstruct the final time-domain signal.

Table 4.1: Evaluation of objective scores in seen SNRs utilizing the TIMIT and DNS challenge dataset.

Dataset	Objective Measures	-5 dB		0 dB		5 dB	
		Noisy signal	Enhanced signal	Noisy signal	Enhanced signal	Noisy signal	Enhanced signal
TIMIT	PESQ	1.25	2.09	1.39	2.62	1.62	3.08
	STOI	0.57	0.78	0.67	0.83	0.77	0.86
	LSD	11.16	2.2	10.95	1.89	11.33	1.7
DNS Challenge	PESQ	1.69	2.23	2.01	2.74	2.42	3.13
	STOI	0.8	0.82	0.87	0.85	0.93	0.87
	LSD	12.86	2.08	12.65	1.99	13.03	1.94

4.3.3 Results

To assess the effectiveness of the proposed network in improving speech quality and intelligibility, we employ various objective metrics. For evaluating intelligibility, we utilize the STOI metric ² [57], which quantifies the clarity and comprehensibility of speech signals, with higher scores indicating better intelligibility. To measure speech quality, we use the PESQ metric ³ [55], which reflects the perceived quality of speech signals, with higher scores indicating superior quality. Additionally, we assess speech distortion using the LSD metric [75], which evaluates the similarity between target and enhanced signals in the spectral domain. Smaller LSD values indicate reduced distortion or improved similarity between signals. These metrics collectively offer a thorough evaluation of the proposed network’s performance in enhancing speech quality and intelligibility.

4.3.3.1 Objective evaluation on RDNet in seen SNRs

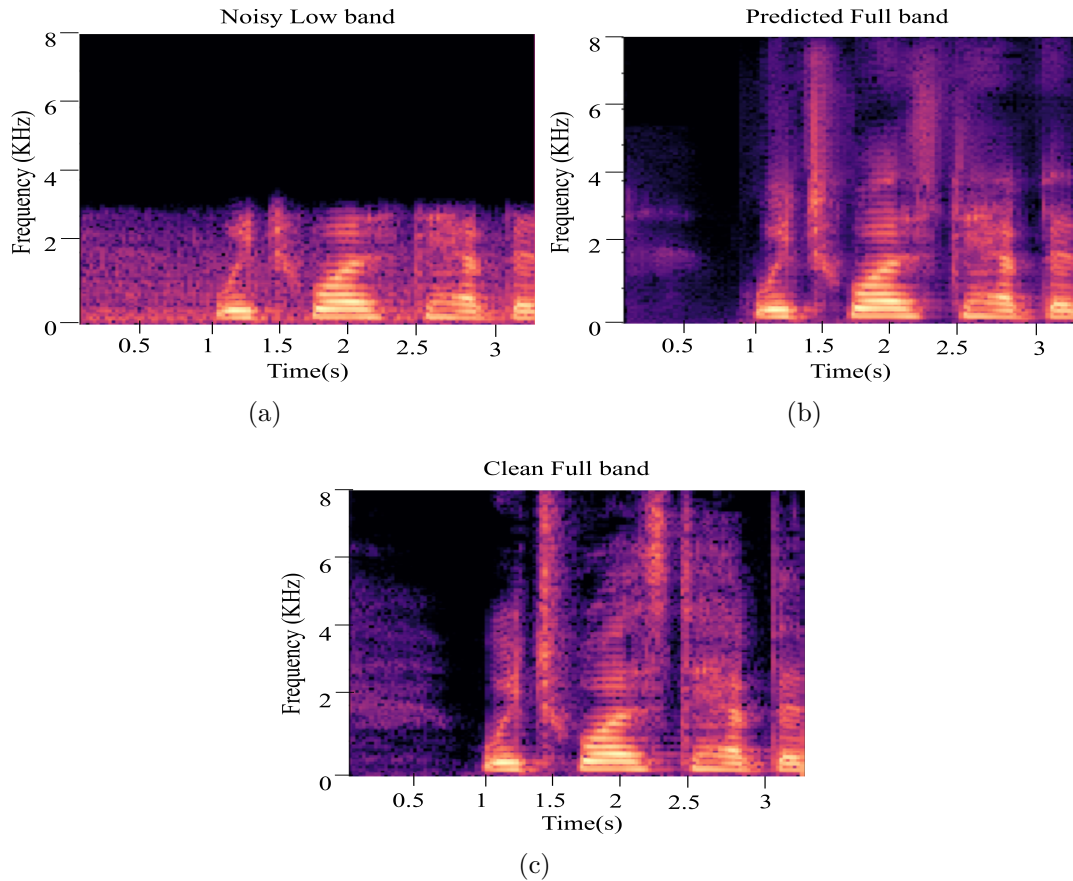
The Table 4.1 provides a comparative analysis of objective measures for noisy and enhanced signals at three different SNR levels (-5 dB, 0 dB, and 5 dB) across two datasets: TIMIT and DNS Challenge in seen noises. In terms of PESQ, the enhanced signals consistently outperform the noisy signals for both datasets at all SNR levels. However, the improvement in PESQ scores varies between the datasets, with DNS Challenge exhibiting higher gains compared to TIMIT. For example, at -5 dB SNR, the PESQ score for DNS Challenge enhanced signals is 2.23, representing an improvement of 0.54 over the noisy signal, whereas for TIMIT, the improvement is 0.84 (from 1.25 to 2.09). Similarly, in terms of STOI, the enhanced signals show improvement over the noisy signals for TIMIT dataset across all SNR levels. DNS Challenge consistently demonstrates higher STOI scores compared to TIMIT, but the improvement over noisy signals is not noteworthy. Finally, LSD scores also exhibit improvement in the enhanced signals for both datasets and across all SNR levels, with DNS Challenge showing slightly

²discussed in section 3.3.3

³discussed in section 3.3.3

Table 4.2: Evaluation of objective scores in Unseen SNRs utilizing the TIMIT and DNS challenge dataset.

Dataset	Objective Measures	-2.5 dB		2.5 dB		7.5 dB	
		Noisy signal	Enhanced signal	Noisy signal	Enhanced signal	Noisy signal	Enhanced signal
DNS Challenge	PESQ	1.69	2.25	2.01	2.66	2.42	3.09
	STOI	0.8	0.8	0.87	0.85	0.93	0.86
	LSD	12.86	3.11	12.65	2.77	13.03	2.42
TIMIT	PESQ	1.25	2.03	1.39	2.54	1.62	3.03
	STOI	0.57	0.7	0.67	0.78	0.77	0.82
	LSD	11.16	2.16	10.95	1.92	11.33	1.8

**Figure 4.5:** Spectrogram of one test utterance with SNR 5 dB, (a) Noisy low band signal spectrogram, (b) Predicted full band signal spectrogram, (c) Clean full band signal spectrogram with oracle phase.

better results than TIMIT. Overall, the enhanced signals from DNS Challenge dataset tend to exhibit better objective measures compared to TIMIT, indicating superior performance in terms of speech quality and intelligibility.

4.3.3.2 Objective evaluation on RDNet in Unseen SNRs

The Table 4.2 presents a comparative analysis of objective measures obtained from two datasets, DNS Challenge and TIMIT, for different signal-to-noise ratio (SNR) levels (-2.5 dB, 2.5 dB, and 7.5 dB). In terms of the PESQ metric, the

enhanced signals from the DNS Challenge dataset consistently outperform those from the TIMIT dataset across all SNR levels, exhibiting notable improvements ranging from 0.44 to 0.75. Specifically, at -2.5 dB SNR, the PESQ score for the DNS Challenge dataset is 2.25 compared to 2.03 for TIMIT, representing a 35.2% improvement. This trend continues across the other SNR levels, with percentage improvements of 37.8% and 36.4% at 2.5 dB and 7.5 dB SNR, respectively. Similarly, for the STOI metric, the enhanced signals from the DNS Challenge dataset consistently yield higher scores compared to TIMIT, with improvements ranging from 0.03 to 0.09. Notably, at -2.5 dB SNR, the STOI score for the DNS Challenge dataset is 0.8 compared to 0.7 for TIMIT. This trend persists across other SNR levels, with percentage improvements of 5.2% and 4.9% at 2.5 dB and 7.5 dB SNR, respectively. Additionally, for the LSD metric, the enhanced signals from the DNS Challenge dataset exhibit substantially lower distortion compared to TIMIT, with improvements ranging from 1.19 to 1.64. The enhanced signals from the DNS Challenge dataset consistently demonstrate superior performance across all evaluated objective measures compared to those from the TIMIT dataset. The DNS Challenge dataset exhibits higher PESQ scores, indicating improved perceived speech quality, higher STOI scores, suggesting enhanced speech intelligibility, and lower LSD scores, reflecting reduced signal distortion. These findings underscore the effectiveness of the DNS Challenge dataset in facilitating the development of robust speech enhancement algorithms.

4.3.3.3 Spectrogram analysis

Figure 4.5 provides a visual representation of a single utterance extracted from the DNS challenge test set. the proposed RDNet, as shown in Figure 4.5 produces cleaner signal at low-frequency and richer high frequency content. Also it can be observed that the enhanced-fullband representations no spectral discontinuity provide wellpresented features for subsequent extension operation.

4.4 Summary

In this chapter, a novel end-to-end frequency-domain framework is designed to simultaneously enhance and extend speech signals. This framework optimizes both a mask-based speech enhancement module and an ideal bandwidth extension module using multi-task learning. The approach involves separating the signals into magnitude and phase spectra. To assess its effectiveness and robustness, we conduct experiments using two distinct datasets. The experimental results demonstrate significant improvements with our proposed method, achiev-

ing a relative enhancement of 36.4% and 62% over the TIMIT dataset in terms of perceptual evaluation of speech quality (PESQ) and log-spectral distortion (LSD), respectively, when compared to the DNS challenge dataset. Additionally, our method boasts a more compact design in terms of the number of parameters employed.

Chapter 5

CSDNet:Conformer Motivated Super Denoised Network

This chapter provides a brief discussion CSDNet: Conformer motivated Super Denoised Network, a deep learning network for joint noise reduction and bandwidth expansion to restore speech signals. By integrating a Dilated Convolutional U-Net model with recurrent and conformer layers, CSDNet effectively reduces noise while expanding bandwidth. Experimental results show superior performance compared to existing models, both objectively and subjectively, across various noise conditions. CSDNet preserves speech intelligibility and quality while reducing noise, making it a promising solution for real-world applications.

Section 5.1 provides insight into the motivation behind the proposed framework and reviews related studies. Section 5.2 outlines the proposed architecture, while Section 5.3 details the experimental setup and Section 5.4 shows performance evaluation across various conditions. Finally, Section 5.5 summarizes the chapter, highlighting the key findings and contributions of the research.

5.1 Motivation

Humans demonstrate an exceptional ability to selectively focus attention on specific speech signals amidst noisy acoustic environments, including scenarios with competing speech, such as bubble noise. Speech bandwidth extension algorithms aim to emulate this selective attention mechanism, filtering out background noise to prioritize the important speech content. These algorithms serve as crucial pre-processing modules in various real-world applications, including automatic speech recognition (ASR), speaker identification, and hearing aids design[76, 77]. Over the years, numerous statistical approaches have been proposed to mimic human selective attention, such as soft-decision noise suppression filters[78], minimum-

mean-square-error (MMSE)-based spectral amplitude estimators[7], wideband coding algorithms[2]. These methods endeavor to estimate optimal multiplicative masking through statistical inference to suppress noise effectively. The emergence of deep neural networks (DNNs) has revolutionized speech bandwidth extension by introducing learning-based models to predict clean full band speech from distorted low band inputs, initially in the frequency domain. While frequency-domain methods have been widely explored, they have limitations, including suboptimal feature extraction using short-time Fourier transform (STFT) and challenges in accurate phase reconstruction, leading to compromised speech quality[79, 80, 81, 82, 83]. To address these limitations, recent advancements have shifted towards end-to-end time-domain approaches, operating directly on the noisy waveform without explicit STFT manipulation. These approaches, such as mask-based noiserobust bandwidth extension framework with multi-task learning (MTL-MBE)[84], optimize the entire encoder-decoder structure and eliminate the need for phase estimation. However, they may exhibit sensitivity to varying testing environments due to their reliance on learned parameters from training data[85, 86]. To mitigate these challenges, for taking the advantages of two domain, this section has attempted to combine frequency-domain and time-domain networks to achieve balanced results. However, the performance of such methods remains suboptimal, inspiring the development of novel approaches. In this chapter, a Conformer motivated super denoised network (CSDNet) is proposed to enhance robustness in varying testing environments for extending the bandwidth of the speech by removing noise. Primary aims of this study are-

- to design a specialized super-denoised network architecture explicitly tailored for the joint tasks of noise reduction and bandwidth expansion from low-resolution spectrograms containing low-frequency components.
- to evaluate the effectiveness of the proposed model across a range of noise conditions and signal-to-noise ratios (SNRs). This evaluation will provide insights into the robustness and generalizability of the model, thereby assessing its suitability for real-world applications in varied environments.
- to conduct a formal subjective study involving human listeners to assess the perceptual quality and intelligibility of the enhanced speech signals. By analyzing the results of this subjective study, we aim to gain a deeper understanding of how the proposed model impacts the subjective user experience and its potential implications for practical deployment.

5.2 Proposed Architecture

In this section, the Conformer Motivated Super Denoised network (CSDNet) is presented, tailored to address the challenge posed by low-band degraded speech signals. The CSDNet is specifically designed to restore lost frequency bands in recovered speech signals, with a focus on handling audio signals characterized by diminished low-frequency components. By incorporating both magnitude and phase information, the architecture aims to effectively reconstruct missing high-frequency bands in speech signals. The utilization of Dilated Convolutional Parallel Recurrent Neural Network components is pivotal in achieving optimal performance in reconstructing degraded signals. The overarching objective of the CSDNet is to enhance the quality of speech signals, thereby contributing to improved intelligibility.

The overall architecture of the CSDNet (depicted in Figure 5.1) introduces a novel neural network structure called Conformer Motivated Super Denoised network (CSDNet) for speech bandwidth extension and restoration. This network simultaneously processes degraded low-frequency components in input speech by adopting a U-Net-style Convolutional Neural Network (CNN) as an encoder-decoder structure. The encoder is designed to receive upsampled low-frequency input spectrograms in both the frequency and time domains, incorporating a T-F transformation layer to mitigate distortions for achieving sub-optimal performance. The input undergoes analysis through convolutional blocks, generating a latent embedding. The encoder and decoder comprise four convolution and deconvolution blocks, leveraging the expansive receptive fields of strided convolution layers to enhance contextual analysis and improve network representation capability. Additionally, a Parallel Recurrent Neural Network (PRNN) and an attention-based conformer layer are integrated between the encoder and decoder to further reinforce sequential modeling. These enhancements address the limitations of sequential modeling within the encoder’s convolutional layers. U-Net skip connections are employed to maintain information flow between the encoder and decoder blocks during network propagation. The final output, an enhanced full-band spectrogram, is obtained by downsampling the decoder output. This comprehensive architecture integrates advanced features, including dilated convolutions, parallel recurrent modeling, attention mechanisms, and U-Net skip connections, resulting in a robust and efficient network design.

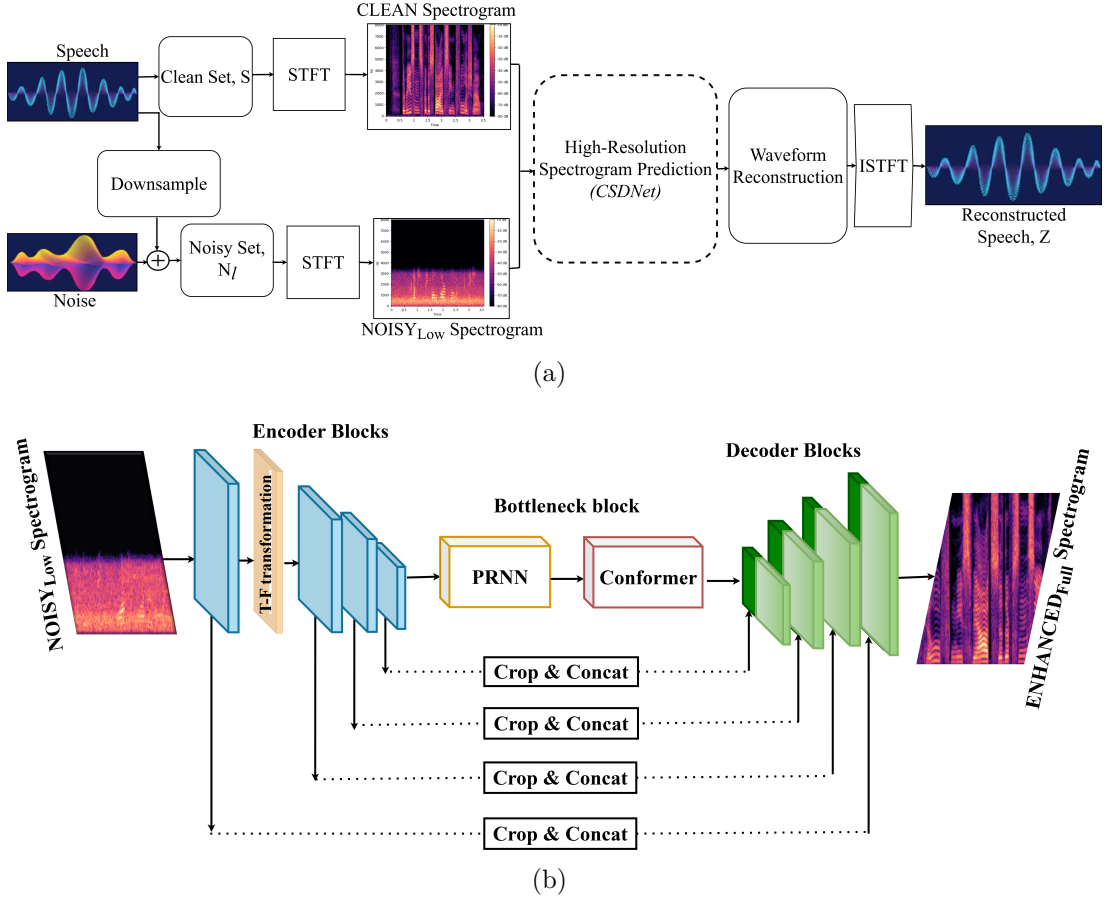


Figure 5.1: An illustration of the proposed network architecture(CSDNet). The processing chain depicted in (a) illustrates the comprehensive prediction core, utilizing the Short-Time Fourier Transform (STFT) signal transformation, high-resolution spectrogram prediction, and the Inverse STFT process. Meanwhile, the modular components illustrated in (b) represent individual cores, each equipped with learned feature transformations based on 2D-Conv encoder-decoder layers, a Time-Frequency (T-F) transformation layer, a Parallel Recurrent Neural Network (PRNN), and a Conformer module.

5.2.1 Encoder-Decoder

In each encoder-decoder block of the network, real-valued convolution layers are utilized to ensure sequential enhancement of both magnitude and phase, facilitating the decoder’s utilization of all spectral and temporal features learned by the encoder. This architecture employs complex convolution to enhance the quality of reconstructed speech signals, distinguishing it from networks using real-valued layers. The following description delves into the algorithmic formulation and practical implementation.

Within the proposed network, the real-valued encoder layer comprises an input layer of dimensions (32, 129), signifying the input divided into 129 frames

with the Short-Time Fourier Transform (STFT) computed for each frame to provide improved frequency resolution. Each magnitude matrix, into which the spectrogram is partitioned, has a size of 32 for enhanced granularity analysis. The convolutional layers consist of 32 filters and operate by sliding a kernel matrix over the input matrix, performing point-wise multiplication. These layers employ the Rectified Linear Unit (ReLU) activation function and adopt 'same' padding to maintain input dimensions. A dilation rate of (1, 2) is introduced to influence the effective receptive field of the filters, with a subsequent application of max pooling with a size of (2, 2) to further reduce spatial dimensions.

The decoder segment begins with an upsampling layer applied to the output of the preceding layer, effectively doubling its spatial dimensions. This is followed by a cropping layer to determine the crop shape necessary for concatenation, establishing a skip connection to minimize the semantic gap between the features from the encoder and decoder blocks. Two transposed convolutional layers, each with 64 filters, ReLU activation, 'same' padding, and a dilation rate of (2, 2), are subsequently implemented.

Mathematically, the decoder input is represented as:

$$\text{Input}(i, j) = \text{concatenation} \left(\begin{array}{l} \text{Upsampling}(D_l), \\ \text{Cropping}(E_l) \end{array} \right) \quad (5.1)$$

where i and j denote the spatial dimensions, and D_l and E_l denote the decoder and encoder layer respectively. The kernel equation for both encoder and decoder convolutional layers is given as $\text{Kernel}(m, n) = \text{Kernel}(a, b)$, where a , b , m , and n range from 1 to 3. The resulting output equation for convolution is expressed as:

$$\text{Output}(x, y) = \text{ReLU} \left(\sum_{a=1}^3 \sum_{b=1}^3 \text{Input}(x - a + 1, y - b + 1) \times \text{Kernel}(a, b) \right) \quad (5.2)$$

5.2.2 T-F Transformation Layer

Inspired by the TFT-Net Framework for speech enhancement [87], this section introduces a T-F transformer layer (shown in Figure 5.2) designed to enhance temporal and frequency representations of spectrogram data. The layer begins with a 1x1 Conv2D operation followed by reshaping along the time and frequency axes, facilitating channel-wise information exchange. Subsequently, Conv1D operations are applied independently along these reshaped axes, enabling cross-axis

interactions to capture local dependencies across spectral and temporal dimensions. The transformed tensors undergo element-wise multiplication with the input spectrogram along both axes, and the resulting outputs are concatenated. A Bidirectional LSTM layer along the time axis captures global dependencies, producing outputs that are reshaped and concatenated with the initial Conv2D layer output. Finally, another 1x1 Conv2D operation yields the ultimate transformed output.

Mathematically, the operation on the input spectrogram $\text{Input}(i, j, k)$ involves a 1x1 Conv2D operation with 32 filters $\text{Kernel}(m, n, l)$. The resulting tensor is reshaped along the time and frequency axes, and Conv1D operations are independently applied along these reshaped axes. The output tensors $\text{Output}_{\text{Conv1D}}(i, j, k)$ are reshaped back to the original shape and multiplied element-wise with the input spectrogram. Concatenation of the outputs $\text{Output}_{\text{Concat}}(i, j, k)$ precedes processing by a Bidirectional LSTM layer along the time axis, resulting in $\text{Output}_{\text{LSTM}}(i, j, k)$. This output is reshaped and concatenated with the initial Conv2D layer output to produce the final output $\text{Output}(x, y, z)$. The equations describing these operations underscore the interplay between time and frequency components, encapsulating the functionality of the T-F transformer layer.

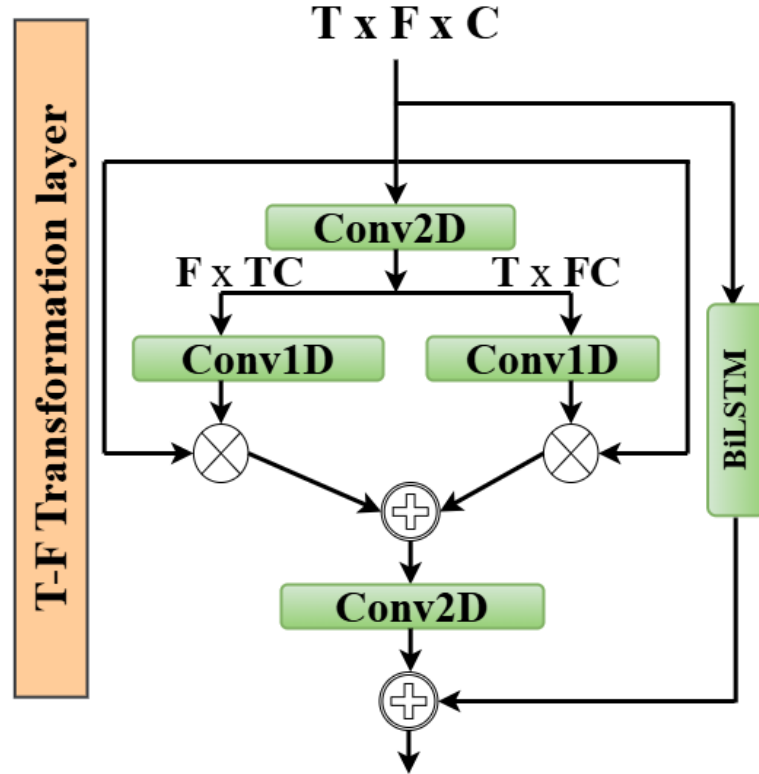


Figure 5.2: Architecture of T-F Transformation Layer.

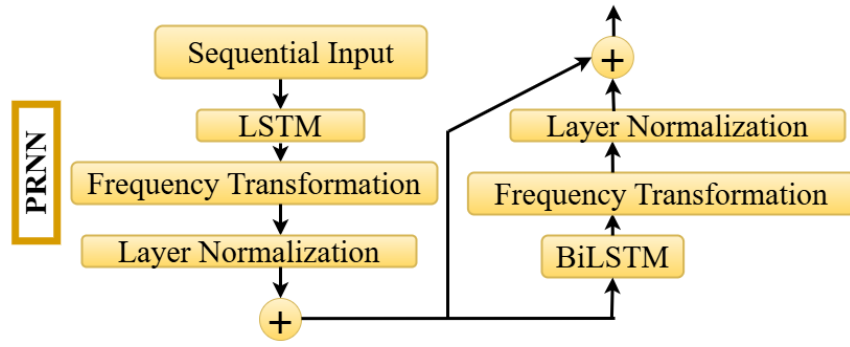


Figure 5.3: Architecture of Parallel RNN(PRNN) visualizing the data dividing into segments, analyzing those segments, and finally reassembling them.

5.2.3 Parallel Recurrent Neural Network (PRNN)

Drawing inspiration from the Dual-Path RNN structure [88] for speech separation, this study integrates the Parallel Recurrent Neural Network (PRNN) architecture, comprising three key components: segmenting the data, analyzing these segments, and reassembling them. Initially, the input is segmented with overlap, forming a 3D tensor. This tensor is then fed into stacked PRNN blocks, which employ both local and global modeling alternately. The output of the last layer is converted back to a sequential format using the overlap-concatenation technique. Figure 5.3 illustrates the functioning of the model.

The PRNN architecture, situated as a bottleneck layer between the encoder and decoder, is designed to leverage both sequential and parallel processing for efficient feature extraction and learning. Initially, the input tensor undergoes reshaping into a 3D tensor. Subsequently, an LSTM layer with 256 units processes sequential information along the temporal axis, capturing long-range dependencies locally. Following the LSTM layer, a frequency transformation layer employs a bias-free Dense layer with a specified activation function and layer normalization to stabilize training. The resulting output is concatenated with the original input tensor, and a subsequent 1D convolutional layer reshapes this concatenated tensor, introducing nonlinear transformations. The architecture also integrates a Bidirectional LSTM layer to capture bidirectional temporal dependencies globally. Another frequency transformation layer, similar to the previous one, follows the Bidirectional LSTM layer. The resulting tensor is normalized, concatenated with the previously reshaped tensor, and passed through a final 1D convolutional layer. This meticulous architecture combines the advantages of parallel and sequential processing, enhancing the network’s capability for feature extraction and learning.

5.2.4 Conformer Block

Within the bottleneck layer of the network architecture, an attention-based conformer block is integrated, playing a pivotal role in both the encoder and decoder. This component is meticulously crafted to capture intricate sequential dependencies following the PRNN block. Drawing inspiration from the Convolution-augmented Transformer proposed by Gulati et al. [72] for speech recognition, the conformer block combines multi-head attention with 4 heads, feedforward neural networks (FFN), and convolutional neural networks (CNN). Specifically, the conformer block in this study consists of two feedforward modules enveloping the Multi-Headed Attention module and the Convolution module. This architectural configuration, reminiscent of the Macaron-Net [73], serves as the fundamental framework. Unlike the original Conformer proposed by Gulati et al. [72], where two Macaron-net style feed-forward layers with half-step residual connections were utilized, our adaptation in the CSDNet model employs two vanilla-style feed-forward layers with residual connections. This setup, with feed-forward layers strategically positioned around the attention and convolution modules, leads to significant improvements, enhancing the block’s capability to model complex sequential relationships. The proposed modification, inspired by the Macaron-Net structure, presents a novel approach to organizing feed-forward layers, thereby contributing to the overall effectiveness of the Conformer block in capturing intricate dependencies within sequential data.

5.2.5 Skip Connection

This subsection explores the U-Net architecture, emphasizing the crucial role of skip connections and the incorporation of crop and concatenation operations between the encoder and decoder. Recent research underscores the importance of facilitating feature exchange between these components. It is observed that the initial encoder layer captures low-level local spectral and temporal features, progressively extracting higher-level features with subsequent layers. Conversely, the final decoder layer receives processed features from its precursor and low-level features from the first encoder layer via a skip connection, potentially encountering feature mismatch issues. To mitigate this, the proposed method involves cropping the previous encoder layer and integrating it within the skip connection, thereby adapting encoder features for improved compatibility with the decoder. These skip connections act as direct channels for high-resolution feature transfer, addressing spatial information loss during downsampling and effectively alleviating the bottleneck problem. This design not only preserves fine-grained spatial

details but also enhances the model’s ability to capture both global and local features. By aligning spatial dimensions and merging features, this approach enhances feature reuse and facilitates gradient flow during backpropagation. This contributes to enhanced model convergence and training stability, making U-Net architectures particularly suitable for tasks such as image segmentation and speech dereverberation.

5.3 Experimental Setup

5.3.1 Speech Corpus

This research aims to assess the effectiveness of the proposed CSDNet model through experiments conducted on the TIMIT database¹ [54]. The TIMIT database consists of 6,300 speech utterances in American English, each accompanied by phonetic transcriptions, with durations ranging from 3 to 5 seconds. To create our training dataset, we selected a subset of 500 utterances from 50 different speakers from the TIMIT database. These sentences were then modified by filtering out high-frequency components and introducing noise from four distinct sources, resulting in the creation of the Noisy Lowpass dataset. This dataset includes three Signal-to-Noise Ratios (SNRs): -5, 0, and 5 dB. Within the training dataset, we set aside 150 randomly selected utterances to form a validation test set. The environmental noise conditions encompass various sources such as airports, babble, cars, and speech-shaped noise (SSN). Notably, airport and car noises are classified as low-band, while babble and SSN belong to the high-band or mid-band category. For model evaluation, we utilized a secondary subset comprising 50 samples for testing. These utterances were mixed with three ”unseen” noise types (Restaurant, Train, and Exhibition) and four ”seen” noise types (Airport, Babble, Car, SSN) across six different SNR levels: -5, -2.5, 0, 2.5, 5, and 7.5 dB. It’s noteworthy that three of the six SNR levels are consistent with those in the training set. The term ”seen” noise refers to noise types encountered during model training, while ”unseen” noise denotes types unfamiliar to the trained model. Overall, the training set consists of 45,000 noisy-clean pairs, while the test set comprises 1,000 pairs. All speech samples and noises were resampled at a rate of 16 kHz.

¹discussed in section 3.3.1

5.3.2 Experimental Framework

The experimental setup initiates with the instantiation of the Adam optimizer, configured with a learning rate set to 0.001. The proposed model is compiled utilizing this optimizer, with the Mean Absolute Error (MAE) designated as the chosen loss function. To efficiently generate batches of training data during model training, custom instances of the Data Generator class are employed. Training extends over 50 epochs, with the number of steps per epoch determined by dividing the length of the input dataset by the batch size. Subsequent to the training phase, model evaluation on the test data is conducted using the evaluate method. This entails computing loss values and metrics on the provided test data, with the resulting metrics displayed for further analysis. For the testing and reconstruction phase, a frame reconstruction function is defined to reshape the model predictions and input test data to the appropriate dimensions. The model’s prediction method is then applied to generate predictions on the test data. The resulting magnitude predictions undergo reconstruction using the defined function. Additionally, reconstruction of the input and target phases is performed to facilitate subsequent computations. Depending on the specified phase (e.g., ‘oracle’, ‘imaged_inverse’, ‘imaged_noisy’, ‘Griffin_Lim’), full phase information is reconstructed accordingly. The predicted magnitude matrix is multiplied by the exponential of the full phase information, followed by application of the inverse short-time Fourier transform (iSTFT) to reconstruct the final time-domain signal.

5.3.3 Evaluation Metrics

To gauge the efficacy of the proposed network concerning speech quality and intelligibility, we utilize a range of objective metrics. For intelligibility assessment, we compute the short-time objective intelligibility (STOI)² [57] and Spectrogram Orthogonal Polynomial Measure (SOPM)³ [56]. These metrics quantify the clarity and comprehensibility of the speech signal, with higher values indicating improved intelligibility. To evaluate speech quality, we employ the Perceptual Evaluation of Speech Quality (PESQ)⁴ [55] and mean opinion score (MOS). PESQ scores, typically ranging from -0.5 to 4.5, reflect the perceived quality of the speech signal, with higher values signifying superior quality. MOS values, computed based on CSIG, CBAK, and COVL scores, range from 1 to 5, representing different aspects of speech quality. Additionally, we assess speech distortion using the

²discussed in section 3.3.3

³discussed in section 3.3.3

⁴discussed in section 3.3.3

Log Spectral Distance (LSD), where smaller values indicate reduced distortion or improved similarity between signals.

- **Log Spectral Distance (LSD):** LSD is a metric used to evaluate the distortion between two signals in the spectral domain. It measures the similarity between the magnitude spectra of the target and enhanced signals on a logarithmic scale. A smaller LSD value indicates a higher similarity or reduced distortion between the signals. The range of LSD typically varies, but smaller values are preferred, indicating better performance [75].
- **CSIG (Conversational Speech Intelligibility Grade):** CSIG is a component of the mean opinion score (MOS) used to evaluate speech quality. It assesses the overall intelligibility of the speech signal in conversational scenarios. CSIG values range from 1 to 5, with higher values indicating better speech intelligibility [89].
- **CBAK (Conversational Background Noise Attenuation):** CBAK is another component of MOS used to evaluate speech quality. It measures the ability of a system to attenuate background noise while preserving speech quality. CBAK values also range from 1 to 5, with higher values indicating better background noise attenuation [89].
- **COVL (Conversational Overall Listening Quality):** COVL is a component of MOS that provides an overall assessment of the listening experience, considering both speech quality and background noise attenuation. Similar to CSIG and CBAK, COVL values range from 1 to 5, with higher values indicating better overall listening quality [89].

These objective metrics provide a comprehensive evaluation of the proposed network’s performance, offering insights into its ability to enhance both speech quality and intelligibility.

5.3.4 Baseline Networks

Here four reference baselines are implemented, each serving a specific purpose related to audio super resolution or bandwidth extension. One of these baselines, referred to as DNN [52], is tailored for ideal bandwidth extension under noisy conditions. The remaining three baselines—CNN Unet [37], TFiLM [38], and AFiLM [39] are designed specifically for audio super-resolution aka bandwidth extension scenarios.

DNN [52]: A 3-layer network that predicts missing high-frequency components from low-frequency signals using a spectral mapping function inspired by [34]. This model operates effectively in both noisy and clean conditions.

CNN [37]: A fully convolutional encoder-decoder framework that maps narrowband signals to wideband in the time domain. To enhance the time dimensions during upscaling, subpixel shuffling layers are incorporated in the up-sampling blocks. The inclusion of skip connections accelerates the training process.

TFiLM [38]: Utilizes a temporal-FiLM (Feature-Wise Modulation) approach to capture Long-Range Sequence Dependencies. This architectural component, inspired by adaptive batch normalization and its extensions, employs a recurrent neural network to modify the activations of a convolutional model. This results in an expanded receptive field for convolutional sequence models with minimal computational overhead.

AFiLM [39]: Employs a Self-Attention based temporal-FiLM, combining convolution and self-attention. This model uses a self-attention mechanism instead of recurrent neural networks to modulate the activations of the convolutional model. Furthermore, it allows for increased parallelization, resulting in significantly faster training compared to traditional approaches.

5.4 Results

5.4.1 Assessment Category

Speech bandwidth extension (SBE) serves to enhance speech signals' quality and comprehensibility, particularly in scenarios where limited bandwidth may hinder transmitting full-frequency speech. Within SBE assessment, two main aspects are considered: speech intelligibility and speech quality. Speech intelligibility assessment focuses on how well a listener can understand the transmitted speech, evaluating clarity and accuracy. It employs objective metrics to gauge a listener's ability to comprehend spoken content accurately, crucial for effective communication in telecommunications, broadcasting, and speech recognition systems. Conversely, speech quality assessment aims to evaluate the overall perceived quality of the speech signal, considering subjective factors like naturalness and listener satisfaction. This assessment relies on subjective listening tests where human listeners provide ratings based on their perception of speech quality. Metrics like Mean Opinion Score (MOS) are commonly used. Speech quality assessment is pertinent in applications like telecommunications, audio streaming, and any context where user experience with the speech signal is vital. In essence, speech intelligibil-

Table 5.1: Results of the metrics for assessing intelligibility and quality of speech are presented in this table, including the training condition. Note that all results are evaluated under the noisy TIMIT testset. Scores were calculated for STOI(0 $\sim$ 1), SOPM(0 $\sim$ 1), PESQ(-0.5 $\sim$ 4.5), CSIG, CBAK, COVL are respectively (1 $\sim$ 5).

SNR (dB)	Method	STOI				SOPM				PESQ			
		Airport	Babble	Car	SSN	Airport	Babble	Car	SSN	Airport	Babble	Car	SSN
-5	Noisy	0.53	0.49	0.49	0.57	0.53	0.53	0.53	0.53	1.3	1.26	1.21	1.25
	Proposed	0.76	0.75	0.75	0.79	0.87	0.88	0.88	0.89	2.07	2.02	2.03	2.25
0	Noisy	0.66	0.63	0.63	0.67	0.7	0.7	0.69	0.7	1.35	1.34	1.32	1.39
	Proposed	0.83	0.82	0.82	0.84	0.9	0.9	0.91	0.91	2.62	2.53	2.58	2.73
5	Noisy	0.79	0.76	0.76	0.77	0.83	0.82	0.83	0.83	1.56	1.54	1.5	1.62
	Proposed	0.86	0.85	0.85	0.87	0.91	0.91	0.91	0.92	3.04	2.97	2.97	3.19
SNR (dB)	Method	CSIG				CBAK				COVL			
		Airport	Babble	Car	SSN	Airport	Babble	Car	SSN	Airport	Babble	Car	SSN
-5	Noisy	1	1	1	1	1.81	1.79	1.82	1.89	1	1	1	1
	Proposed	3.13	2.95	3.05	3.23	2.92	2.87	2.91	3.05	2.73	2.61	2.65	2.85
0	Noisy	1	1	1	1	1.97	1.93	1.98	2.01	1	1	1	1
	Proposed	3.48	3.37	3.53	3.57	3.14	3.14	3.17	3.17	3.08	3	3.12	3.14
5	Noisy	1	1	1	1	2.19	2.16	2.2	2.29	1	1	1	1
	Proposed	3.92	3.82	3.87	4.31	3.41	3.33	3.4	3.5	3.53	3.4	3.49	3.8

ity assessment focuses on understanding speech content accurately, while speech quality assessment considers the overall subjective experience, including naturalness and pleasantness. Evaluating the proposed SBE technique involves assessing its performance in handling both noisy and degraded speech, demonstrating satisfactory results in various conditions. Objective measurements assessing speech intelligibility and quality are analyzed and compared with established networks like DNN, CNN, TFiLM, and AFiLM across different noise conditions and signal-to-noise ratios (SNRs) in both familiar and unfamiliar scenarios.

5.4.2 Comparative Objective Evaluation of Speech Intelligibility and Quality Metrics

Table 5.1 provides a detailed comparison of objective scores, assessing the performance of the proposed CSDNet method against the baseline Noisy method across varying Signal-to-Noise Ratios (SNRs) and noise types. The evaluation includes essential metrics such as Speech Transmission Index (STOI), Speech Quality Objective Performance Measure (SOPM), Perceptual Evaluation of Speech Quality (PESQ), as well as additional quality metrics (CSIG, CBAK, and COVL). Scores are presented for both unprocessed noisy signals and signals enhanced by the CSDNet method, covering four distinct noise types—Airport, Babble, Car, and SSN—across an SNR range from -5 dB to 5 dB. Each score represents the average intelligibility or quality derived from 50 utterances. The results demonstrate the consistent superiority of the CSDNet method over the Noisy method across all metrics and noise conditions, indicating its effectiveness in improving

Table 5.2: Quantitative evaluation of proposed Speech Bandwidth Extension algorithm (CSDNet) for six SNR groups in both seen and unseen noises.

SNRs (dB)	STOI				SOPM				PESQ				LSD			
	Seen Noises		Unseen Noises		Seen Noises		Unseen Noises		Seen Noises		Unseen Noises		Seen Noises		Unseen Noises	
	NoisyLow	Enh. Full	NoisyLow	Enh. Full	NoisyLow	Enh. Full	NoisyLow	Enh. Full	NoisyLow	Enh. Full	NoisyLow	Enh. Full	NoisyLow	Enh. Full	NoisyLow	Enh. Full
-5	0.52	0.76	0.54	0.69	0.53	0.88	0.51	0.59	1.25	2.09	1.21	1.83	10.76	2.43	10.83	3.33
-2.5	0.58	0.79	0.56	0.70	0.61	0.89	0.59	0.66	1.28	2.28	1.21	1.85	10.70	2.19	10.70	2.95
0	0.64	0.82	0.66	0.72	0.70	0.90	0.69	0.78	1.35	2.61	1.32	1.911	10.81	1.94	10.88	2.76
2.5	0.71	0.83	0.66	0.76	0.77	0.91	0.76	0.81	1.44	2.77	1.32	2.08	10.70	1.74	11.01	2.4
5	0.77	0.86	0.72	0.8	0.83	0.91	0.83	0.88	1.55	3.13	1.50	2.72	11.24	1.66	11.52	2.03
7.5	0.81	0.86	0.78	0.81	0.87	0.91	0.87	0.88	1.69	3.19	1.65	3.03	11.5	1.48	11.67	1.65
Mean	0.67	0.82	0.65	0.75	0.72	0.91	0.712	0.77	1.43	2.68	1.37	2.24	10.95	1.90	11.10	2.52

both speech intelligibility and quality. Particularly noteworthy is the CSDNet method’s remarkable performance in enhancing speech intelligibility, especially evident in car noise scenarios compared to babble and SSN noise. Notably, the CSDNet method achieves a substantial improvement of +65% in speech quality assessment for SSN noise conditions.

5.4.3 Evaluation of CSDNet Performance Across Various Noise Conditions and SNRs

Table 5.2 presents the quantitative evaluation of the CSDNet algorithm’s performance across different noise conditions and signal-to-noise ratios (SNRs), both in seen and unseen scenarios. The objective is to assess the algorithm’s ability to enhance speech intelligibility and quality under varying conditions. Objective metrics including STOI, SOPM, PESQ, and LSD are computed for different SNR groups ranging from -5 dB to 7.5 dB for seen and unseen noise types. The results demonstrate that the CSDNet consistently outperforms noisy low-band speech, showing significant improvements in speech intelligibility and quality across all noise types and SNRs. Notably, the algorithm exhibits higher relative enhancements in seen noise conditions compared to unseen ones. It achieves competitive speech quality scores even in unfamiliar noise environments while maintaining high levels of speech intelligibility. The objective scores for enhanced full-band speech consistently exceed those for noisy low-band speech, with greater improvements observed in STOI and SOPM metrics as SNRs increase. Overall, there is a notable reduction in log spectral distance for both seen and unseen noise conditions.

5.4.4 Impact of Phase on Reconstructed Speech

Figure 5.4 depicts the influence of different phases (Oracle, Imaged_noisy, Imaged_Inverse, and Griffin Lim) on objective measurements, offering a comprehensive assessment of the CSDNet method compared to the baseline Noisy method. The bars, color-coded in orange, yellow, and green, correspond to SNRs of -

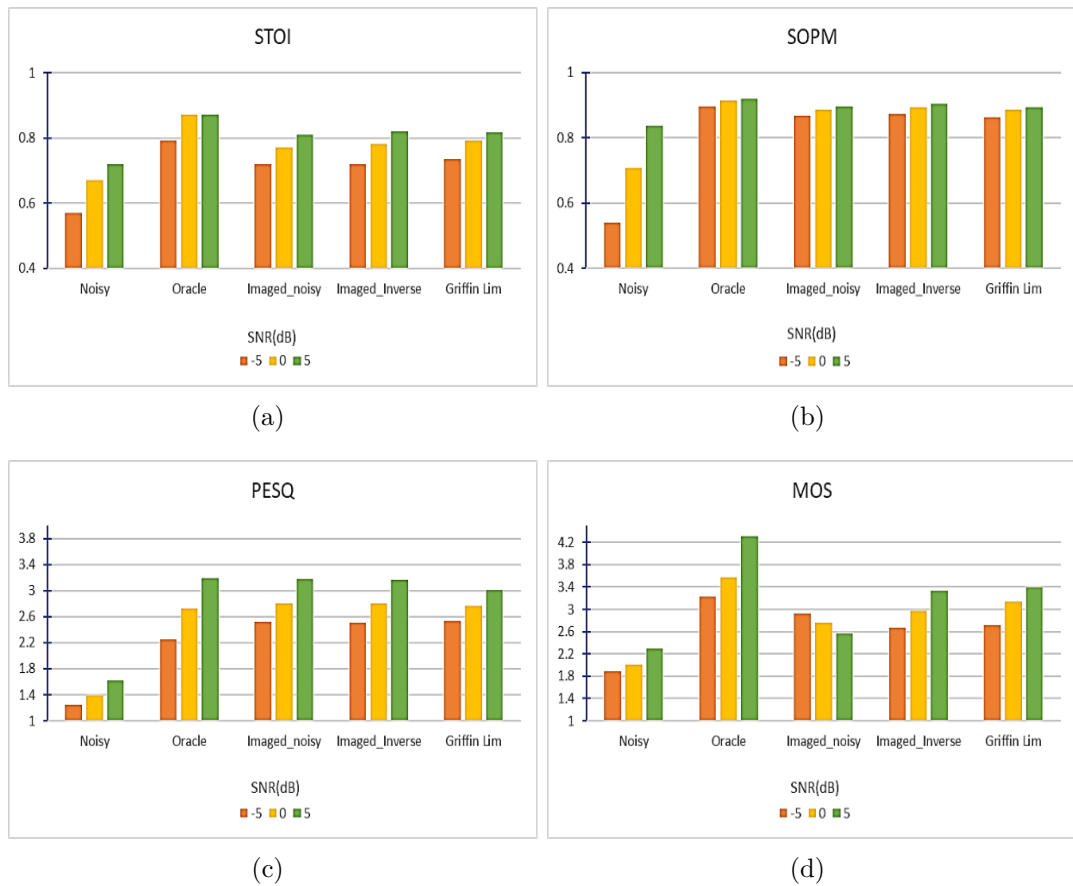


Figure 5.4: Objective measurements on various phase of input test sets of proposed models. Oracle, Imaged_noisy, Imaged_inverse and Griffin Lim, respectively are four different phases for the reconstruction of test sets. Orange, Yellow, and Green bars represent the SNRs of -5 dB, 0dB and 5 dB, respectively.

5 dB, 0 dB, and 5 dB, respectively. This analysis examines how variations in phase affect reconstructed speech, providing insights into the performance of the proposed models across diverse conditions. Objective metrics, including STOI, SOPM, PESQ, and additional quality metrics (CSIG, CBAK, COVL), are evaluated across three noise types—SSN, Airport, Babble, and Car—across an SNR range from -5 dB to 5 dB. Regarding speech intelligibility metrics (STOI and SOPM), the CSDNet consistently outperforms the Noisy method across all SNRs and noise types. Notably, there is a significant improvement in STOI and SOPM at an SNR of -5 dB, indicating enhanced speech intelligibility for Babble and Car noises. The Oracle phase serves as a benchmark for optimal performance. For speech quality metrics (PESQ, CSIG, CBAK, COVL), the CSDNet demonstrates superior performance across SNRs and noise types. Particularly, there is a notable increase in PESQ scores for all noise types at an SNR of 0 dB, highlighting the method’s effectiveness in enhancing overall speech quality. The comprehen-

sive evaluation consistently emphasizes the advantage of the CSDNet over the Noisy method in both speech intelligibility and quality metrics, showcasing its robustness in challenging acoustic environments. These results underscore the potential of the CSDNet as a valuable enhancement for speech signals, especially in diverse noise conditions and SNRs.

The equations for signal reconstruction in different phases are as follows:

1. Oracle Phase:

$$Z_k(n) = \text{ISTFT}(|Z_k(n)| \cdot \exp(1.0j \cdot \angle Y_k(n))) \quad (5.3)$$

where $\angle Y_k(n)$ represents the phase angle of the Clean Full band spectrogram at frequency bin k and frame n .

2. Imaged_Noisy Phase:

$$Z_k(n) = \text{ISTFT}(|Z_k(n)| \cdot \exp(1.0j \cdot \angle X_k(n))) \quad (5.4)$$

where $\angle X_k(n)$ represents the phase angle of the Noisy Low band spectrogram at frequency bin k and frame n .

3. Imaged_Inverse Phase:

$$Z_k(n) = \text{ISTFT}(|Z_k(n)| \cdot \exp(1.0j \times \text{np.concatenate}((\angle X_k(n), -\text{flip}(\angle X_k(n))))) \quad (5.5)$$

where $\angle X_k(n)$ is the phase angle of the Noisy Low band spectrogram at frequency bin k and frame n , and $-\text{flip}(\angle X_k(n))$ is the phase angle of the negative flipped Noisy Low band spectrogram at frequency bin k and frame n , used as the high-band phase in the reconstructed signal.

4. Griffin Lim Phase:

$$Z_k(n) = \text{ISTFT} \left(|Z_k(n)| \times \frac{\sum_{m=1}^M X_m(n-1) \cdot \exp(1.0j \cdot \angle Y_k(n-1))}{\sum_{m=1}^M |X_m(n-1)| \cdot \exp(1.0j \cdot \angle Y_k(n-1))} \right) \quad (5.6)$$

where $Z_k(n)$ represents the complex value of the spectrogram of the reconstructed signal at frequency bin k and frame n , $|Z_k(n)|$ denotes the magnitude of the reconstructed spectrogram at frequency bin k and frame n , and M is the total number of frequency bins.

Table 5.3: Evaluation of Model Components through Ablation Study.

Model	STOI	SOPM	MOS	PESQ	LSD
Unet	0.76	0.861	2.845	2.947	2.069
Unet+T-F layer	0.794	0.896	3.088	2.38	2.089
Unet+T-F layer+PRNN	0.8	0.902	3.22	2.53	1.92
Proposed CSDNet	0.87	0.903	3.73	3.19	1.68

5.4.5 Investigating the Component Effectiveness in Proposed Model through Ablation Studies

Table 5.3 presents the results of an ablation study conducted to assess the impact of individual components on the performance of the proposed CSDNet model. Four variations of the model were created to isolate the effects of specific components: 'Unet'— Utilizing only the typical Dilated Convolutional Unet architecture without additional components., 'Unet+T-F layer'—including the T-F transformation layer in the encoder block, 'Unet+T-F layer+PRNN'—including the parallel RNN block in the bottleneck layer, and 'Proposed CSDNet'—including the Unet and T-F layer in encoder and PRNN and conformer block in the bottleneck layer. The proposed 'CSDNet' serves as the baseline for comparison. Results indicate the following: 'Unet' performs slightly lower across most metrics compared to the full model, emphasizing the significance of the excluded components. Adding blocks with Unet shows improvements in STOI and SOPM, highlighting the positive impact of these components on speech intelligibility. Notably, the CSDNet without PRNN in bottleneck exhibits a decrease in performance, emphasizing the crucial role of the parallel RNN block in maintaining overall quality. The comprehensive study provides valuable insights into the contributions of individual components to the effectiveness of the proposed CSDNet model in enhancing overall performance. The proposed CSDNet demonstrates superior performance across all metrics compared to the ablated models, highlighting the significance of the T-F transformer, PRNN, and Conformer components in enhancing speech intelligibility and quality.

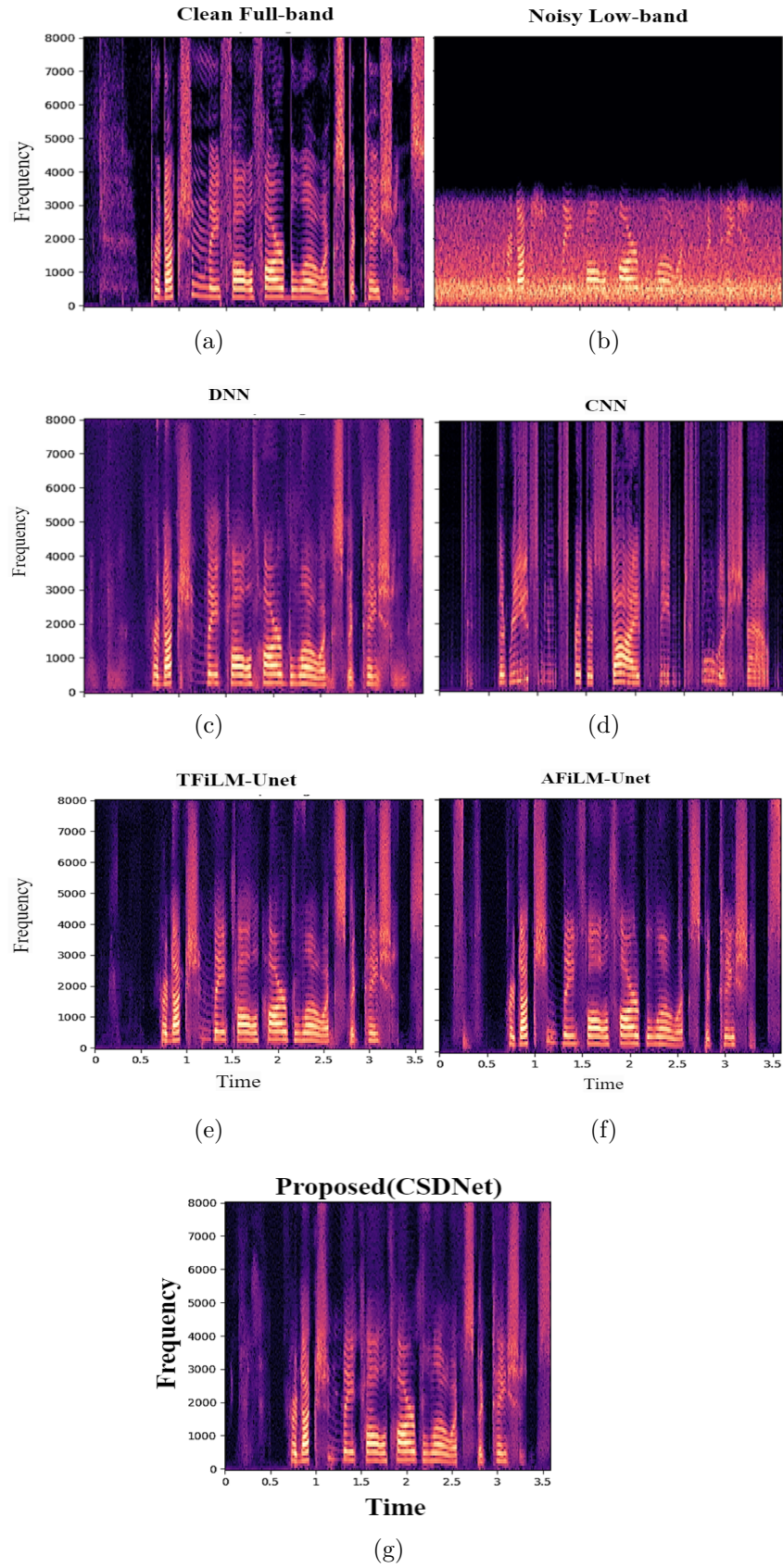


Figure 5.5: Qualitative results. Spectrogram of (a) clean full band speech, (b) noisy low band speech, (c) DNN output, (d) CNN output, (e) TFiLM-Unet output, (f) AFiLM-Unet output and (g) CSDNet output (Proposed).

Table 5.4: A Comparative result of all metrics among different existing models is presented in this table. The upper arrow represents that higher is better, and vice versa. Note that all results are evaluated under both clean and noisy TIMIT test sets. The approaches are performed in both frequency domain and time domain.

Models	CLEAN Condition						NOISY Condition					
	STOI $\uparrow$	CSIG $\uparrow$	CBAK $\uparrow$	COVL $\uparrow$	PESQ $\uparrow$	LSD $\downarrow$	STOI $\uparrow$	CSIG $\uparrow$	CBAK $\uparrow$	COVL $\uparrow$	PESQ $\uparrow$	LSD $\downarrow$
DNN [52]	0.87	2.45	3.08	2.19	3.08	1.22	0.84	2.27	3.06	2.17	1.62	2.72
CNN [37]	0.854	2.07	3.38	2.65	3.11	2.01	0.83	2.33	3.42	2.25	2.99	3.08
TFiLM-Unet [38]	0.93	2.27	3.25	2.17	2.23	2.72	0.78	2.45	3.23	2.09	1.79	2.8
AFiLM-Unet [39]	0.94	2.64	3.21	2.46	2.55	2.29	0.72	1.18	3.45	1.38	1.74	2.97
Proposed	0.92	3.76	3.2	3.47	4.09	1.08	0.86	3.52	3.13	2.96	3.01	1.79

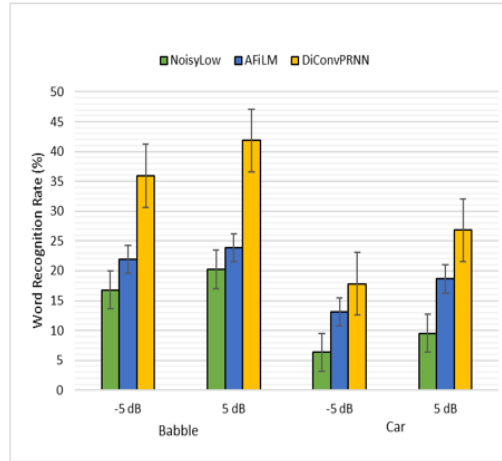
5.4.6 Comparative Analysis

1. **Comparative Assessment against Existing Networks:** When juxtaposed with established models like DNN, CNN, TFiLM, and AFiLM, primarily engineered for clean speech enhancement and bandwidth extension, CSDNet model, as depicted in Table 5.4, showcases remarkable superiority across various metrics such as PESQ, CSIG, CBAK, COVL, STOI, and LSD. Notably, traditional Speech Bandwidth Extension (SBE) models were traditionally trained solely with clean speech data, constraining their adaptability to noisy conditions. Our model demonstrates exceptional performance, particularly in metrics highly sensitive to noise like CSIG (3.52) and COVL (2.96). In contrast to conventional BWE models (DNN, CNN, TFiLM, AFiLM), CSDNet exhibits significant advancements, surpassing them across multiple metrics. In comparison to state-of-the-art models such as TFiLM and AFiLM, our proposed network showcases improved performance in STOI, PESQ, and notably, a substantial reduction in LSD (-1.18). Moreover, it’s worth noting that while achieving competitive speech quality metrics, CSDNet emerges as a lightweight model. The proposed network achieves approximately a 32% reduction (68.22M $\rightarrow$ 36.5M) and an impressive 91% reduction (127.43M $\rightarrow$ 36.5M) in model parameters compared to TFiLM and AFiLM, respectively. This underscores the efficiency and effectiveness of the model in delivering enhanced performance with reduced computational complexity.
2. **Spectrogram Comparison:** To illustrate the disparities in the recovered enhanced-plus-extended signal between baseline methods and our proposed CSDNet, a specific speech sample from the test set is utilized, as illustrated in Figure 5.5. The examination reveals that the CSDNet (Figure 5.5(g)) yields a cleaner signal at low frequencies and richer high-frequency content compared to DNN (Figure 5.5(c)). Spectrograms of the clean, noisy,

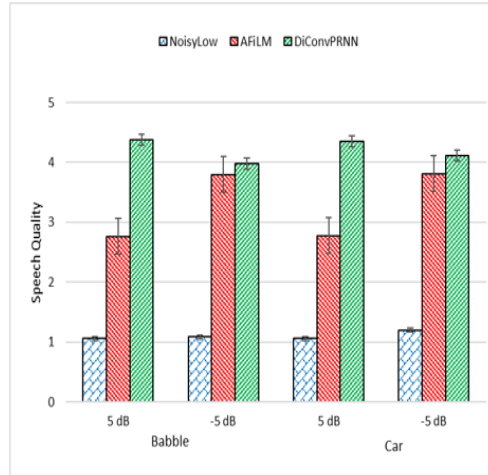
baseline-processed, and proposed network-processed signals are presented to assess the extended enhanced output. The results demonstrate that our proposed network effectively mitigates noise while preserving harmonic speech structure in the spectrogram. While DNN and CNN-based methods exhibit some capability to recover high-frequency components, the DNN-based recovery tends to wash away the spectral structure, whereas the CNN-based method struggles to learn it well, leading to noise introduction in the final expanded speech. Comparing our proposed method with TFiLM and AFiLM-based approaches, the harmonic structures in the expanded speech are notably clearer in the high-frequency component, particularly during specific time intervals. In TFiLM, the harmonics in the spectrum of the reference method appear blurred, indicating distortion introduced by the mere use of the mirror version phase in expansion. Conversely, our expanded speech exhibits better-recovered structure, closely resembling the original one. The high clarity of the harmonics in the spectrogram underscores the superiority of estimating the recovered spectrogram.

5.4.7 Comparative Subjective Evaluation of Speech Intelligibility and Quality Metrics

The speech intelligibility and quality metrics were assessed under babble and car noise conditions at -5 dB and 5 dB SNR for eight NH recipients. Error bars, indicating ± 1 standard deviation, accompany the results. Under babble noise at -5 dB SNR, the word recognition rate (WRR) scores saw notable enhancements with AFiLM and CSDNet processing. Specifically, WRR scores improved from 16% to 21.36% with AFiLM and to 37.39% with CSDNet. Similarly, at 5 dB SNR, mean WRR scores improved significantly from 22.90% to 42.83%. In the presence of car noise, CSDNet exhibited improved performance over AFiLM across both -5 dB and 5 dB SNR conditions. The improvement was by 4% and 10.5%, respectively. Regarding speech quality, under babble noise at 5 dB SNR, quality scores showed enhancements from 1.06% to 2.76% with AFiLM processing and to 4.27% with CSDNet. At -5 dB SNR, mean quality scores improved from 3.7% to 3.98%. In the presence of car noise, CSDNet again outperformed AFiLM, showing improvements by 0.4% and 2.5% at -5 dB and 5 dB SNR, respectively. Figure 5.6(a) illustrates the speech intelligibility metrics, depicting the significant improvements achieved with DiConvPRNN processing under both babble and car noise conditions. On the other hand, Figure 5.6(b) showcases the speech quality metrics, highlighting the superior performance of CSDNet over AFiLM across



(a)



(b)

Figure 5.6: Subjective Speech Enhancement Performance based on (a) Speech Intelligibility, (b) Speech Quality

various SNR levels and noise types.

5.5 Summary

This chapter introduces a novel approach called the Conformer motivated Super Denoised network (CSDNet) to mitigate the mismatch problem arising from sensitive domain-specific encoders/decoders. The proposed method involves embedding a conv Unet with T-F transformation layers in the encoder, PRNN, and

Conformer block at the bottleneck to reduce dependency on training data. Experimental results demonstrate that CSDNet surpasses recent baseline methods in terms of various metrics such as PESQ, LSD, SOPM, MOS, and STOI. Additionally, CSDNet exhibits favorable performance in subjective studies compared to existing methods like AFiLM, indicating its suitability for real-time applications.

Chapter 6

Conclusion

In this thesis, we addressed the challenge of improving speech quality and intelligibility through speech bandwidth extension, focusing on the impact of unseen noises, channel effects, and the sensitivity of time-domain encoders/decoders. Three novel approaches were proposed to enhance the performance of state-of-the-art speech bandwidth extension techniques.

In Chapter 3, we presented a novel deep neural network (DNN)-based approach for simultaneous bandwidth expansion and speech restoration. Our approach aimed to enhance speech signals' quality while increasing their bandwidth and reducing noise interference. Through a comparative evaluation of audio signal processing, we demonstrated the efficacy of the enhanced signal compared to the original noisy low-band signal. Notably, the reconstructed signal with the "oracle" phase exhibited superior performance, achieving a PESQ score of 3.04 and SOPM score of 0.86. These findings underscored the potential of our proposed method to effectively enhance speech signal quality, particularly in noisy environments.

Furthermore, in Chapter 4, we introduced an end-to-end frequency-domain framework designed to simultaneously enhance and extend speech signals. This framework optimized both a mask-based speech enhancement module and an ideal bandwidth extension module using multi-task learning. Experimental results demonstrated significant improvements with our proposed method, achieving a relative enhancement of 36.4% and 62% over the TIMIT dataset in terms of perceptual evaluation of speech quality (PESQ) and log-spectral distortion (LSD), respectively, compared to the DNS challenge dataset. Additionally, our method boasted a more compact design in terms of the number of parameters employed.

Lastly, in Chapter 5, we presented the Conformer motivated Super Denoised network (CSDNet) to address the mismatch problem arising from sensitive domain-

specific encoders/decoders. Experimental results showed that CSDNet outperformed recent baseline methods across various metrics such as PESQ, LSD, SOPM, MOS, and STOI. Moreover, CSDNet exhibited favorable performance in subjective studies compared to existing methods like AFiLM, demonstrating its suitability for real-time applications.

Overall, the findings of this thesis underscored the importance of addressing the mismatch problem in speech enhancement and bandwidth extension tasks. Our proposed approaches offer promising solutions to improve speech signal quality and robustness, paving the way for advancements in various real-world applications. Further research is warranted to explore additional refinements and validate the effectiveness of these methods across diverse datasets and environments.

6.1 Future Scope

Speech bandwidth extension and restoration hold promise for enhancing real-world applications in noisy environments. Our future research will expand upon current efforts in the following areas:

- **Addressing Mismatch Problems:** While our thesis has demonstrated significant improvements in speech enhancement through adversarial training, the focus has primarily been on mitigating the effects of unseen noise. However, real-world scenarios present additional challenges, such as variations in accents and languages, which can contribute to training-testing mismatches. To address this, we aim to enhance speech enhancement methods by incorporating adversarial training to learn more generalized acoustic features capable of disentangling accent and language information.
- **Bandwidth Extension Pre-processing:** In simulating signals corrupted by channel effects, conventional approaches employ a clean-cut low-pass filter to remove high-frequency information, resulting in a clear divide between low and high-frequency signals. However, real-world scenarios need such distinct boundaries, leading to mismatches between training and testing. To overcome this, we propose improving speech enhancement methods through multi-task training with different low-pass filters for pre-processing training data. This approach aims to align training data with real-world scenarios better, thus reducing the mismatch problem.
- **Correct Phase Reconstruction:** Additionally, we recognize the importance of accurate phase reconstruction in speech enhancement. Future re-

search will focus on refining phase reconstruction techniques to ensure the fidelity of enhanced speech signals, further enhancing overall speech quality and intelligibility.

By pursuing these avenues of research, we aim to advance the state-of-the-art in speech bandwidth extension and restoration, making significant strides towards enabling robust and effective communication in challenging real-world environments.

6.2 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.

1. **T. Taher**, N. Mamun, and M. A. Hossain, “A joint bandwidth expansion and speech enhancement approach using deep neural network,” in 2023 International Conference on Electrical, Computer and Communication Engineering (ECCE), pp. 1–4, IEEE, 2023

Bibliography

- [1] N. Hou, “Mismatch problem in deep-learning based speech enhancement,” 2022.
- [2] P. Gajjar, N. Bhatt, and Y. Kosta, “Artificial bandwidth extension of speech & its applications in wireless communication systems: A review,” in *2012 International Conference on Communication Systems and Network Technologies*, pp. 563–568, IEEE, 2012.
- [3] M. Croll, *Sound-quality improvement of broadcast telephone calls*. Research Department, Engineering Division, BBC, 1972.
- [4] J. Zeng and L. Yang, “Speech enhancement of complex convolutional recurrent network with attention,” *Circuits, Systems, and Signal Processing*, vol. 42, no. 3, pp. 1834–1847, 2023.
- [5] 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.
- [6] P. C. Loizou, A. Lobo, and Y. Hu, “Subspace algorithms for noise reduction in cochlear implants,” *The Journal of the Acoustical Society of America*, vol. 118, no. 5, pp. 2791–2793, 2005.
- [7] Y. Ephraim and D. Malah, “Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator,” *IEEE Transactions on acoustics, speech, and signal processing*, vol. 32, no. 6, pp. 1109–1121, 1984.
- [8] I. Cohen and B. Berdugo, “Noise estimation by minima controlled recursive averaging for robust speech enhancement,” *IEEE signal processing letters*, vol. 9, no. 1, pp. 12–15, 2002.
- [9] P. C. Loizou, *Speech enhancement: theory and practice*. CRC press, 2007.
- [10] V. Mitra, H. Franco, R. M. Stern, J. Van Hout, L. Ferrer, M. Graciarana, W. Wang, D. Vergyri, A. Alwan, and J. H. Hansen, “Robust features in deep-learning-based speech recognition,” *New Era for Robust Speech Recognition: Exploiting Deep Learning*, pp. 187–217, 2017.
- [11] S. Pascual, A. Bonafonte, and J. Serra, “Segan: Speech enhancement generative adversarial network,” *arXiv preprint arXiv:1703.09452*, 2017.
- [12] D. Stoller, S. Ewert, and S. Dixon, “Wave-u-net: A multi-scale neural network for end-to-end audio source separation,” *arXiv preprint arXiv:1806.03185*, 2018.
- [13] Y. Luo and N. Mesgarani, “Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation,” *IEEE/ACM transactions on audio, speech, and language processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

- [14] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Joint optimization of masks and deep recurrent neural networks for monaural source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 12, pp. 2136–2147, 2015.
- [15] L. Rabiner, "Digital processing of speech signals," *Prentice Hall google schola*, vol. 2, pp. 601–604, 1978.
- [16] A. V. Oppenheim, *Discrete-time signal processing*. Pearson Education India, 1999.
- [17] M. Bosi and R. E. Goldberg, *Introduction to digital audio coding and standards*, vol. 721. Springer Science & Business Media, 2002.
- [18] M. Kahrs and K. Brandenburg, *Applications of digital signal processing to audio and acoustics*. Springer Science & Business Media, 1998.
- [19] K. K. Paliwal and B. S. Atal, "Efficient vector quantization of lpc parameters at 24 bits/frame," *IEEE transactions on speech and audio processing*, vol. 1, no. 1, pp. 3–14, 1993.
- [20] J. Allen, "Short term spectral analysis, synthesis, and modification by discrete fourier transform," *IEEE transactions on acoustics, speech, and signal processing*, vol. 25, no. 3, pp. 235–238, 1977.
- [21] V. Britanak, "Spectral band replication (sbr) compression technology: Survey of the unified efficient low-cost implementations of complex exponential-and cosine-modulated qmf banks," *Signal Processing*, vol. 115, pp. 49–65, 2015.
- [22] Y. M. Cheng, D. O'Shaughnessy, and P. Mermelstein, "Statistical recovery of wideband speech from narrowband speech," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 544–548, 1994.
- [23] Y. Yoshida, "An algorithm to reconstruct wideband speech from narrowband speech based on code book mapping," in *International Conference on Spoken Language Processing, ICSLP'94*, 1994.
- [24] J. Epps and W. H. Holmes, "A new technique for wideband enhancement of coded narrowband speech," in *1999 IEEE Workshop on Speech Coding Proceedings. Model, Coders, and Error Criteria (Cat. No. 99EX351)*, pp. 174–176, IEEE, 1999.
- [25] S. S. Stevens, J. Volkmann, and E. B. Newman, "A scale for the measurement of the psychological magnitude pitch," *The journal of the acoustical society of america*, vol. 8, no. 3, pp. 185–190, 1937.
- [26] N. Enbom and W. B. Kleijn, "Bandwidth expansion of speech based on vector quantization of the mel frequency cepstral coefficients," in *1999 IEEE Workshop on Speech Coding Proceedings. Model, Coders, and Error Criteria (Cat. No. 99EX351)*, pp. 171–173, IEEE, 1999.
- [27] Y. Agiomyrgiannakis and Y. Stylianou, "Conditional vector quantization for speech coding," *IEEE transactions on audio, speech, and language processing*, vol. 15, no. 2, pp. 377–386, 2007.

- [28] K.-Y. Park and H. S. Kim, “Narrowband to wideband conversion of speech using gmm based transformation,” in *2000 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 00CH37100)*, vol. 3, pp. 1843–1846, IEEE, 2000.
- [29] P. Jax and P. Vary, “Artificial bandwidth extension of speech signals using mmse estimation based on a hidden markov model,” in *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP’03).*, vol. 1, pp. I–I, IEEE, 2003.
- [30] P. Jax and P. Vary, “On artificial bandwidth extension of telephone speech,” *Signal Processing*, vol. 83, no. 8, pp. 1707–1719, 2003.
- [31] M. Nilsson, H. Gustafson, S. V. Andersen, and W. B. Kleijn, “Gaussian mixture model based mutual information estimation between frequency bands in speech,” in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. I–525, IEEE, 2002.
- [32] B. Iser and G. Schmidt, “Neural networks versus codebooks in an application for bandwidth extension of speech signals,” in *INTERSPEECH*, pp. 565–568, 2003.
- [33] J. Kontio, L. Laaksonen, and P. Alku, “Neural network-based artificial bandwidth expansion of speech,” *IEEE transactions on audio, speech, and language processing*, vol. 15, no. 3, pp. 873–881, 2007.
- [34] K. Li, Z. Huang, Y. Xu, and C.-H. Lee, “Dnn-based speech bandwidth expansion and its application to adding high-frequency missing features for automatic speech recognition of narrowband speech,” in *INTERSPEECH*, pp. 2578–2582, 2015.
- [35] X. Liu and C.-C. Bao, “Audio bandwidth extension using ensemble of recurrent neural networks,” *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2016, pp. 1–12, 2016.
- [36] P. B. Bachhav, M. Todisco, M. Mossi, C. Beaugeant, and N. Evans, “Artificial bandwidth extension using the constant q transform,” in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5550–5554, IEEE, 2017.
- [37] V. Kuleshov, S. Z. Enam, and S. Ermon, “Audio super resolution using neural networks,” *arXiv preprint arXiv:1708.00853*, 2017.
- [38] S. Birnbaum, V. Kuleshov, Z. Enam, P. W. W. Koh, and S. Ermon, “Temporal film: Capturing long-range sequence dependencies with feature-wise modulations,” *Advances in Neural Information Processing Systems*, vol. 32, 2019.
- [39] N. C. Rakotonirina, “Self-attention for audio super-resolution,” in *2021 IEEE 31st International Workshop on Machine Learning for Signal Processing (MLSP)*, pp. 1–6, IEEE, 2021.
- [40] M. Mandel, O. Tal, and Y. Adi, “Aero: Audio super resolution in the spectral domain,” in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, IEEE, 2023.
- [41] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, “Wavenet: A generative model for raw audio,” *arXiv preprint arXiv:1609.03499*, 2016.

- [42] S. Mehri, K. Kumar, I. Gulrajani, R. Kumar, S. Jain, J. Sotelo, A. Courville, and Y. Bengio, “Samplernn: An unconditional end-to-end neural audio generation model,” *arXiv preprint arXiv:1612.07837*, 2016.
- [43] C. Dong, C. C. Loy, K. He, and X. Tang, “Image super-resolution using deep convolutional networks,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 2, pp. 295–307, 2015.
- [44] M. Wang, Z. Wu, S. Kang, X. Wu, J. Jia, D. Su, D. Yu, and H. Meng, “Speech super-resolution using parallel wavenet,” in *2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pp. 260–264, IEEE, 2018.
- [45] T. Y. Lim, R. A. Yeh, Y. Xu, M. N. Do, and M. Hasegawa-Johnson, “Time-frequency networks for audio super-resolution,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 646–650, IEEE, 2018.
- [46] I. Vasilakis, “Bandwidth extension for audio signals using deep learning,” 2022.
- [47] A. Pandey and D. Wang, “A new framework for cnn-based speech enhancement in the time domain,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 7, pp. 1179–1188, 2019.
- [48] V. Johnson and I. Parberry, “Music upscaling using convolutional neural networks,” in *Proceedings of the 2020 3rd International Conference on Sensors, Signal and Image Processing*, pp. 58–62, 2020.
- [49] Y. Dong, Y. Li, X. Li, S. Xu, D. Wang, Z. Zhang, and S. Xiong, “A time-frequency network with channel attention and non-local modules for artificial bandwidth extension,” in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6954–6958, IEEE, 2020.
- [50] B. Si, D. Luo, and J. Zhu, “Multi-scale audio super resolution via deep pyramid wavelet convolutional neural network,” *Electronics Letters*, vol. 57, no. 13, pp. 520–522, 2021.
- [51] M. Lagrange and F. Gontier, “Bandwidth extension of musical audio signals with no side information using dilated convolutional neural networks,” in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 801–805, IEEE, 2020.
- [52] T. Taher, N. Mamun, and M. A. Hossain, “A joint bandwidth expansion and speech enhancement approach using deep neural network,” in *2023 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, pp. 1–4, IEEE, 2023.
- [53] J. Makhoul and M. Berouti, “High-frequency regeneration in speech coding systems,” in *ICASSP’79. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 4, pp. 428–431, IEEE, 1979.
- [54] J. S. Garofolo, “Timit acoustic phonetic continuous speech corpus,” *Linguistic Data Consortium, 1993*, 1993.
- [55] I. Rec, “P. 862.2: Wideband extension to recommendation p. 862 for the assessment of wideband telephone networks and speech codecs,” *International Telecommunication Union, CH-Geneva*, vol. 41, pp. 48–60, 2005.

- [56] 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.
- [57] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, “An algorithm for intelligibility prediction of time–frequency weighted noisy speech,” *IEEE Transactions on audio, speech, and language processing*, vol. 19, no. 7, pp. 2125–2136, 2011.
- [58] C. Liu, Q.-J. Fu, and S. S. Narayanan, “Effect of bandwidth extension to telephone speech recognition in cochlear implant users,” *The Journal of the Acoustical Society of America*, vol. 125, no. 2, pp. EL77–EL83, 2009.
- [59] P. S. Nidadavolu, C.-I. Lai, J. Villalba, and N. Dehak, “Investigation on bandwidth extension for speaker recognition.,” in *INTERSPEECH*, pp. 1111–1115, 2018.
- [60] D. Haws and X. Cui, “CycleGAN bandwidth extension acoustic modeling for automatic speech recognition,” in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6780–6784, IEEE, 2019.
- [61] R. Kaminishi, H. Miyamoto, S. Shiota, and H. Kiya, “Investigation on blind bandwidth extension with a non-linear function and its evaluation of x-vector-based speaker verification.,” in *INTERSPEECH*, pp. 4055–4059, 2019.
- [62] H. Yamamoto, K. A. Lee, K. Okabe, and T. Koshinaka, “Speaker augmentation and bandwidth extension for deep speaker embedding.,” in *Interspeech*, pp. 406–410, 2019.
- [63] J. Abel and T. Fingscheidt, “Artificial speech bandwidth extension using deep neural networks for wideband spectral envelope estimation,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 1, pp. 71–83, 2017.
- [64] Y. Gu and Z.-H. Ling, “Waveform modeling using stacked dilated convolutional neural networks for speech bandwidth extension.,” in *INTERSPEECH*, pp. 1123–1127, 2017.
- [65] S. Kim and V. Sathe, “Bandwidth extension on raw audio via generative adversarial networks,” *arXiv preprint arXiv:1903.09027*, 2019.
- [66] A. Gupta, B. Shillingford, Y. Assael, and T. C. Walters, “Speech bandwidth extension with wavenet,” in *2019 IEEE workshop on applications of signal processing to audio and acoustics (WASPAA)*, pp. 205–208, IEEE, 2019.
- [67] P. J. Moreno, B. Raj, and R. M. Stern, “A vector taylor series approach for environment-independent speech recognition,” in *1996 IEEE international conference on acoustics, speech, and signal processing conference proceedings*, vol. 2, pp. 733–736, IEEE, 1996.
- [68] M. L. Seltzer, A. Acero, and J. Droppo, “Robust bandwidth extension of noise-corrupted narrowband speech.,” in *INTERSPEECH*, pp. 1509–1512, 2005.
- [69] H. Seo, H.-G. Kang, and F. Soong, “A maximum a posteriori-based reconstruction approach to speech bandwidth expansion in noise,” in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6087–6091, IEEE, 2014.

- [70] B. Liu, J. Tao, and Y. Zheng, “A novel unified framework for speech enhancement and bandwidth extension based on jointly trained neural networks,” in *2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pp. 11–15, IEEE, 2018.
- [71] Y. Luo, C. Han, N. Mesgarani, E. Ceolini, and S.-C. Liu, “Fasnet: Low-latency adaptive beamforming for multi-microphone audio processing,” in *2019 IEEE automatic speech recognition and understanding workshop (ASRU)*, pp. 260–267, IEEE, 2019.
- [72] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, *et al.*, “Conformer: Convolution-augmented transformer for speech recognition,” *arXiv preprint arXiv:2005.08100*, 2020.
- [73] Y. Lu, Z. Li, D. He, Z. Sun, B. Dong, T. Qin, L. Wang, and T.-Y. Liu, “Understanding and improving transformer from a multi-particle dynamic system point of view,” *arXiv preprint arXiv:1906.02762*, 2019.
- [74] C. K. Reddy, H. Dubey, K. Koishida, A. Nair, V. Gopal, R. Cutler, S. Braun, H. Gamper, R. Aichner, and S. Srinivasan, “Interspeech 2021 deep noise suppression challenge,” *arXiv preprint arXiv:2101.01902*, 2021.
- [75] L. Rabiner, “Fundamentals of speech recognition,” *Prentice Hall google schola*, vol. 2, pp. 447–453, 1993.
- [76] D. Yu, L. Deng, J. Droppo, J. Wu, Y. Gong, and A. Acero, “A minimum-mean-square-error noise reduction algorithm on mel-frequency cepstra for robust speech recognition,” in *2008 IEEE international conference on acoustics, speech and signal processing*, pp. 4041–4044, IEEE, 2008.
- [77] L.-P. Yang and Q.-J. Fu, “Spectral subtraction-based speech enhancement for cochlear implant patients in background noise,” *The journal of the Acoustical Society of America*, vol. 117, no. 3, pp. 1001–1004, 2005.
- [78] R. McAulay and M. Malpass, “Speech enhancement using a soft-decision noise suppression filter,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 2, pp. 137–145, 1980.
- [79] J. Le Roux, H. Kameoka, N. Ono, and S. Sagayama, “Fast signal reconstruction from magnitude stft spectrogram based on spectrogram consistency,” in *Proc. DAFx*, vol. 10, pp. 397–403, 2010.
- [80] N. Sturmel, L. Daudet, *et al.*, “Signal reconstruction from stft magnitude: A state of the art,” in *International conference on digital audio effects (DAFx)*, pp. 375–386, 2011.
- [81] K. Li, B. Wu, and C.-H. Lee, “An iterative phase recovery framework with phase mask for spectral mapping with an application to speech enhancement.,” in *INTERSPEECH*, pp. 3773–3777, 2016.
- [82] N. Zheng and X.-L. Zhang, “Phase-aware speech enhancement based on deep neural networks,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 1, pp. 63–76, 2018.

- [83] K. Yatabe, Y. Masuyama, and Y. Oikawa, “Rectified linear unit can assist griffinlim phase recovery,” in *2018 16th international workshop on acoustic signal enhancement (IWAENC)*, pp. 555–559, IEEE, 2018.
- [84] N. Hou, C. Xu, J. T. Zhou, E. S. Chng, and H. Li, “Multi-task learning for end-to-end noise-robust bandwidth extension,” 2020.
- [85] T. H. Dat, K. Takeda, and F. Itakura, “Generalized gamma modeling of speech and its online estimation for speech enhancement,” in *Proceedings.(ICASSP’05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, vol. 4, pp. iv–181, IEEE, 2005.
- [86] G. Wichern, J. Antognini, M. Flynn, L. R. Zhu, E. McQuinn, D. Crow, E. Manilow, and J. L. Roux, “Wham!: Extending speech separation to noisy environments,” *arXiv preprint arXiv:1907.01160*, 2019.
- [87] C. Tang, C. Luo, Z. Zhao, W. Xie, and W. Zeng, “Joint time-frequency and time domain learning for speech enhancement,” in *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pp. 3816–3822, 2021.
- [88] Y. Luo, Z. Chen, and T. Yoshioka, “Dual-path rnn: efficient long sequence modeling for time-domain single-channel speech separation,” in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 46–50, IEEE, 2020.
- [89] Y. Hu and P. C. Loizou, “Evaluation of objective quality measures for speech enhancement,” *IEEE Transactions on audio, speech, and language processing*, vol. 16, no. 1, pp. 229–238, 2007.