TESIS DE LA UNIVERSIDAD DE ZARAGOZA
2024
479

Jorge Llombart Gil

Progressive Speech Enhancement with Deep Neural Networks

Director/es
Miguel Artiaga, Antonio

ISSN 2254-7606

Prensas de la Universidad
Universidad Zaragoza

Universidad de Zaragoza
Servicio de Publicaciones

ISSN 2254-7606

Universidad Zaragoza
1542

Tesis Doctoral

PROGRESSIVE SPEECH ENHANCEMENT WITH DEEP NEURAL NETWORKS

Autor

Jorge Llombart Gil

Director/es

Miguel Artiaga, Antonio

UNIVERSIDAD DE ZARAGOZA
Escuela de Doctorado

Programa de Doctorado en Tecnologías de la Información y Comunicaciones en Redes Móviles

2024

Universidad Zaragoza
1542

Tesis Doctoral

Progressive Speech Enhancement with Deep Neural Networks

Autor

Jorge Llombart Gil

Director/es

Antonio Miguel Artiaga

ESCUELA DE INGENIERÍA Y ARQUITECTURA
UNIVERSIDAD DE ZARAGOZA
Escuela de Doctorado
Programa de Doctorado en Tecnologías de la Información y Comunicaciones en Redes Móviles

2024

UNIVERSITY OF ZARAGOZA

DOCTORAL THESIS

Progressive Speech Enhancement with Deep Neural Networks

*Author:*Jorge LLombart *Supervisor:*Dr. Antonio Miguel

A thesis submitted in fulfillment of the requirements for the degree of Doctor of Philosophy in Information Technologies and Mobile Network Communications

in the

ViVoLab Research Group
Electronic Engineering and Communications Department

July 23, 2024

Agradecimientos

El camino que he recorrido durante el desarrollo de esta tésis ha sido largo. En este tiempo han habido algunos momentos muy duros, pero la mayoría de los que me vienen a la memoria son inmejorables. Ha sido todo un placer y sobre todo un honor poder colaborar con unas personas tan exelentes tanto en lo personal como en lo profesional. Estas lineas las dedico a todos aquellos que me han acompañado en este camino, que no solo ha sido de crecimiento laboral sino personal.

En primer lugar, quiero darle las gracias a mi director de tesis, Antonio Miguel Artiaga, por todo el apoyo que me ha dado en el desarrollo de esta tesis. Él fué la persona que me mostró un camino dentro de la tecnología, que cuando me encontraba sin guia ni objetivo, me desveló lo increible que sería lo que posteriormente se convertiría en la base de toda mi carrera profesional. Gracias por tu tiempo, gracias por ese entusiasmo contagioso que sirve de luz cuando el camino estaba oscuro, gracias por las ganas de enseñar que siempre me has demostrado, y sobre todo gracias por la confianza que has depositado en mi.

Me gustaría también a gradecer a todos los miembros del grupo ViVoLab, que más que crear un entorno de trabajo acogedor crean una familia para todos. A mis compañeros de laboratorio que además de colaborar y ayudar en todo, siempre estaban dispustos a amenizar las horas de trabajo. Han sido muchos los compañeros a los que estoy agradecido, a Eduardo y Alfonso, un ejemplo a seguir, a Diego, David, Jesús y Paola, que me acogieron cuando llegué, a Julia a Dayana, más que unas compañeras y a Pablo y Victoria que me acompañaron al final.

A mi familia, a mi padre que aunque no está me sigue apoyando, a mi madre que se sigue preocupando. Gracias por guiarme y darme los valores que aunque a veces se me olviden ha hecho que llegue hasta aquí.

A mi mujer, Adriana, que me sostiene, que cree en mí y hace de mi vida una vida feliz. Y por supuesto, a mis hijos que aunque no me dejan trabajar lo hacen con todo el cariño y amor de sus corazoncitos.

A todos los que lo habeís hecho posible, gracias.

Abstract

Speech enhancement is an important field in signal processing, aiming to improve the clarity and intelligibility of speech in noisy environments. This research is crucial for applications like phone calls, hearing aids, and voice-controlled systems, where clear communication is essential. However, existing methods often struggle with complex and varying noise conditions, leading to reduced speech quality and intelligibility. Recent advancements in deep neural networks, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), have shown significant improvements in handling these challenges. These networks can learn intricate patterns in speech and noise, offering more robust solutions. However, they still have limitations, including high computational costs and difficulty in adapting to diverse noise environments. The primary objective of this research is to develop a novel speech enhancement method using Wide Residual Networks (WRNs). This new approach aims to outperform current techniques by providing better speech quality and intelligibility while balancing computational complexity. The goal is to create a more effective speech enhancement procedure that can be integrated into various systems, ensuring clearer and more natural communication in diverse and noisy environments. The proposed architecture processes log magnitude spectrograms, enhancing speech quality through a series of convolutional layers and residual blocks. Experimental results demonstrate that WRNs significantly outperform existing methods, such as RNN-LSTM-based Weighted Prediction Error (WPE), especially in far-field reverberated speech across various room sizes.

Bellow, this work focuses on improving the interpretability of deep learning models used for speech enhancement. Traditional neural network methods often act as "black boxes," making it difficult to understand how they process and enhance speech signals. This issue is addressed by introducing innovative architectures and techniques to visualize and interpret the enhancement process. The thesis presents the Constant Channel Residual Network (CCRN) and the Constant Channel Residual Network with State Path (CCRN-State). These architectures aim to improve speech quality while maintaining a clear

understanding of the internal processes. Progressive Supervision is introduced as a technique to monitor the enhancement process at each network block. This method ensures incremental improvements in speech quality and helps identify critical stages that significantly impact the final output. Experimental results show that these methods not only enhance speech quality but also provide valuable insights into the network's internal mechanisms, leading to a better balance between performance and interpretability. Integrating visualization techniques into deep learning architectures can significantly enhance both the interpretability and effectiveness of speech enhancement models.

Finally the thesis explores the development and application of progressive loss strategies to enhance speech quality through deep learning. The approach involves using Progressive Speech Enhancement (PSE) methods, which improve speech clarity by incrementally refining the enhancement process. The research introduces two main architectures: Progressive Convolutional Neural Networks (P-CNN) and Progressive Residual Networks (P-ResNet). These architectures use novel loss functions—Weighted Progressive (WP) and Uniform Progressive (UP)—to systematically reduce noise and reverberation. Experimental evaluations demonstrate that PSE methods outperform traditional approaches, particularly in noisy and reverberant environments.

This thesis highlights the effectiveness of progressive strategies in stabilizing the training process, ensuring robust performance across different conditions, and setting a new benchmark for speech enhancement technologies.

Resumen

El realce de voz es un campo muy importante en el procesamiento de señales, que busca mejorar la claridad y la claridad del habla en entornos ruidosos. Esta investigación es crucial para aplicaciones como llamadas telefónicas, audífonos y sistemas controlados por voz, donde una comunicación de calidad es esencial. Sin embargo, los métodos existentes a menudo tienen dificultades con condiciones de ruido complejas y variables, lo que lleva a una reducción de la calidad y la inteligibilidad del habla. Los avances recientes en redes neuronales profundas, como las redes convolucionales y las redes neuronales recurrentes, han demostrado mejoras significativas en el manejo de estos desafíos. Estas redes pueden aprender patrones en el habla y el ruido, ofreciendo soluciones más robustas. No obstante, aún presentan limitaciones, incluyendo alto coste computacional y dificultades para adaptarse a entornos de ruido diversos. El objetivo principal de esta investigación es desarrollar un nuevo método de mejora del habla utilizando redes residuales anchas o *Wide Residual Networks*. Este nuevo enfoque pretende superar las técnicas actuales proporcionando una mejor calidad e inteligibilidad del habla, equilibrando la complejidad computacional. El objetivo es crear un procedimiento de mejora del habla más efectivo que pueda integrarse en cualquier sistema, asegurando una comunicación más nítida y natural en entornos diversos y ruidosos. La arquitectura propuesta procesa el logaritmo de la magnitude del espectrograma, mejorando la calidad del habla a través de una serie de capas convolucionales y bloques residuales. Los resultados experimentales demuestran que las WRNs superan significativamente a los métodos existentes, como el Error de Predicción Ponderado (WPE) basado en RNN-LSTM, especialmente en el habla reverberada en campo lejano a través de varios tamaños de habitación.

Esta tesis además se centra en mejorar la interpretabilidad de los modelos de aprendizaje profundo utilizados para la mejora del habla. Los métodos tradicionales de redes neuronales a menudo actúan como "cajas negras," lo que dificulta entender cómo procesan y mejoran las señales de habla. Este problema se aborda introduciendo arquitecturas y técnicas innovadoras para visualizar e interpretar el proceso de

mejora. La tesis presenta las redes residuales con canales constantes (CCRN) y las redes residuales con canales constantes y camino de estado (CCRN-State). Estas arquitecturas tienen como objetivo mejorar la calidad del habla manteniendo una comprensión clara de los procesos de la red. Se introduce la Supervisión Progresiva como una técnica para monitorear el proceso de mejora en cada bloque de la red. Este método asegura mejoras incrementales en la calidad del habla y ayuda a identificar etapas críticas que impactan significativamente el resultado final. Los resultados experimentales muestran que estos métodos no solo mejoran la calidad del habla, sino que también proporcionan valiosas perspectivas sobre el funcionamiento interno de la red, llevando a un mejor equilibrio entre rendimiento e interpretabilidad. Integrar técnicas de visualización en arquitecturas de aprendizaje profundo puede mejorar significativamente tanto la interpretabilidad como la efectividad de los modelos de mejora del habla.

Por último la tesis explora el desarrollo y la aplicación de estrategias de pérdida progresiva para mejorar la calidad del habla a través del aprendizaje profundo. El enfoque implica el uso de métodos de Mejora Progresiva del Habla (PSE), que mejoran la claridad del habla refinando incrementalmente el proceso de mejora. La investigación introduce dos arquitecturas principales: redes convolucionales progresivas (P-CNN) y redes residuales progresivas (P-ResNet). Estas arquitecturas utilizan funciones de coste novedosas como la progresión ponderada (WP) y la progresión uniforme (UP) para reducir sistemáticamente el ruido y la reverberación. Las evaluaciones experimentales demuestran que los métodos PSE superan a los enfoques tradicionales, particularmente en entornos ruidosos y reverberantes.

La tesis destaca la efectividad de las estrategias progresivas en la estabilización del proceso de entrenamiento, asegurando un rendimiento robusto en diferentes condiciones y estableciendo un nuevo estándar para las tecnologías de mejora del habla.

Contents

List of Figures

List of Tables

List of Abbreviations

AMR **A**daptive **M**ulti-**R**ate
ASR **A**utomatic **S**peech **R**ecognition
BWE **B**and**W**idth **E**xpansion
CCRN **C**onstant **C**hannel **R**esidual **N**etwork
CNN **C**onvolutional **N**eural **N**etwork
CNN-WRN **C**onvolutional **N**eural **N**etwork - **W**ide **R**esidual **N**etwork
DAE **D**enoising **A**uto-**E**ncoder
DCT **D**iscrete **C**osine **T**ransform
DNNS **D**eep **N**eural **N**etworks
DNN **D**eep **N**eural **N**etwork
FB **F**ilter **B**anks
FFT **F**ast **F**ourier **T**ransform
GAN **G**enerative **A**dversarial **N**etwork
GMM **G**aussian **M**ixture **M**odel
GRU **G**ated **R**ecurrent **U**nit
HMM **H**idden **M**arkov **M**odel
IBM **I**deal **B**inary **M**ask
IRM **I**deal **R**atio **M**ask
ITU **I**nternational **T**elecommunication **U**nion
LAH **L**ist **A**bbreviations **H**ere
LLR **L**og-**L**ikelihood **R**atio
LPC **L**inear **P**redictive **C**oding
LPCNet **L**inear **P**redictive **C**oding **Net**work
LSA **L**og-**S**pectral **A**mplitude
LSTM **L**ong **S**hort-**T**erm **M**emory
MAE **M**ean **A**bsolute **E**rror
MCD **M**el-**C**epstral **D**istortion

MEL **M**el **F**requency **C**epstrum
MFCC **M**el-**F**requency **C**epstral **C**oefficients
MMLU **M**easured **M**el-**F**requency **C**epstrum
MMSE **M**inimum **M**ean **S**quare **E**rror
MMSE-LSA **M**inimum **M**ean **S**quare **E**rror-**L**og **S**pectral **A**mplitude
MMSE-STSA **M**inimum **M**ean **S**quare **E**rror-**S**hort **T**ime **S**pectral **A**mplitude
MOS **M**ean **O**pinion **S**core
MSE **M**ean **S**quared **E**rror
NIST **N**ational **I**nstitute of **S**tandards and **T**echnology
NMF **N**on-negative **M**atrix **F**actorization
NNE **N**eural **N**etwork **E**nhancement
PESQ **P**erceptual **E**valuation of **S**peech **Q**uality
PLP **P**erceptual **L**inear **P**redictive
POLQA **P**erceptual **O**bjective **L**istening **Q**uality **A**ssessment
PReLU **P**arametric **R**ectified **L**inear **U**nit
P-CNN **P**rogressive **C**onvolutional **N**eural **N**etwork
P-ResNet **P**rogressive **R**esidual **N**etwork
PSE **P**rogressive **S**peech **E**nhancement
RASTA **R**elative **A**pectral **T**ransform **A**mplitude
RASTI **R**apid **S**peech **T**ransmission **I**ndex
REVERB **Re**alistic **Re**verberation
RIR **R**oom **I**mpulse **R**esponse
RNN **R**ecurrent **N**eural **N**etwork
RNNoise **R**ecurrent **N**eural **N**etwork **Noise** Suppression
RSD **R**andom **S**ingular **D**ecomposition
SDR **S**ignal-to-**D**istortion **R**atio
SE **S**peech **E**nhancement
SEGAN **S**peech **E**nhancement **G**enerative **A**dversarial **N**etwork
SNR **S**ignal-to-**N**oise **R**atio
SPE **S**peech **P**rocessing **E**valuation
SRMR **S**peech-to-**R**everberation **M**odulation **E**nergy **R**atio
STFT **S**hort-**T**ime **F**ourier **T**ransform
STOI **S**hort-**T**ime **O**bjective **I**ntelligibility
TFM **T**ime-**F**requency **M**asking
TIMIT **T**exas **I**nstruments **M**assachusetts **I**nstitute of **T**echnology
TNRD **T**rained **N**onlinear **R**eaction-**D**iffusion
TTS **T**ext-**T**o-**S**peech
UP **U**niform **P**rogressive

VAD **V**oice **A**ctivity **D**etection
VAE **V**ariational **A**uto**E**ncoder
WGAN **W**asserstein **G**enerative **A**dversarial **N**etwork
WP **W**eighted **P**rogressive
WPE **W**eighted **P**rediction **E**rror
WRN **W**ide **R**esidual **N**etwork

Dedicated To my Family

Chapter 1

Introduction

1.1 Motivation and context

Communication is one of the most essential aptitudes that define human beings. It is through communication that we are able to express our thoughts, share knowledge, build relationships, and achieve personal and collective progress. Among the various modes of communication, speech holds a main role as it allows for the direct and nuanced exchange of ideas and emotions. Speech is not just a tool for conveying information; it is a fundamental aspect of human interaction.

In the model of communication, there are three key components: the sender, the message, and the receiver. The sender (or emitter) is the person who articulates the message, while the receiver is the one who interprets and understands it. This exchange occurs within an environment that can either facilitate or hinder effective communication. A favorable environment enhances clarity and comprehension, enabling the message to be conveyed accurately. Conversely, an unfavorable environment, characterized by noise and other distortions, can obstruct the communication process, leading to misunderstandings and loss of information.

The necessity for effective speech communication becomes even more pronounced in environments where background noise is prevalent. Such environments include busy urban areas, workplaces with machinery, and crowded public spaces. In these settings, the presence of

noise can significantly degrade the quality of speech, making it difficult for the receiver to accurately interpret the message. This is where speech enhancement becomes crucial. By employing advanced technologies and methodologies, speech enhancement aims to mitigate the adverse effects of noise and other distortions, ensuring that the communication process remains clear and effective. This not only aids in better understanding but also in maintaining the naturalness of speech, which is vital for meaningful human interactions.

In today's world, technology plays a crucial role in enhancing our communication. There are several scenarios where speech enhancement is particularly beneficial because alternative communication aids are unavailable. For instance, in phone calls, we cannot see our interlocutor, and non-verbal cues are absent. This makes clear and precise speech even more important.

Since the COVID-19 pandemic, our reliance on phone calls and video conferences has significantly increased. We now often interact with colleagues, friends, and family through these mediums. Companies are very interested in improving the quality of these communications to provide better remote experiences. This is not only important in convenient settings like home offices or conference rooms but also in more challenging environments. These include crowded offices, busy cafes, or even during conference calls from moving vehicles.

Moreover, the challenge extends to integrating these speech enhancement systems into embedded devices like cell phones, automotive multimedia systems, and even motorcycle communication devices. Ensuring high-quality speech in such varied and often noisy environments requires sophisticated technology. Effective speech enhancement can transform these experiences, making conversations clearer and more natural, regardless of the surroundings.

The push to embed these technologies in everyday devices highlights their growing importance. As we continue to adapt to new ways of interacting remotely, the demand for effective speech enhancement will only increase. By improving speech clarity in diverse environments, we can ensure better communication, fostering more productive and meaningful interactions across various platforms and scenarios.

The research work presented in this thesis was conducted within the Voice Input Voice Output Laboratory (ViVoLab) research group, part of the Aragón Institute for Engineering Research (I3A) at the University of Zaragoza. This work was performed under the supervision of Dr. Antonio Miguel Artiaga.

1.2 Thesis objectives

This thesis focuses on improving speech enhancement using neural networks. The key idea is to develop new methods that enhance speech quality more effectively than existing neural network techniques. We aim to create a speech enhancement procedure that not only outperforms current methods but also provides a balance between computational complexity and performance. This balance is crucial for integrating these methods into a wide range of systems, from simple devices to more complex applications.

Additionally, we aim to understand what happens within the neural network during speech enhancement. To achieve this, we focus on block-distributed architectures, allowing us to examine each block and gain insights into the processes occurring within the network. This approach helps us optimize performance and provides a clearer understanding of the network's behavior.

The key technology selected for this thesis is the Wide Residual Neural Network (WRN). The fundamental idea behind WRNs is their block-distributed architecture, which divides the network into manageable sections or blocks. This approach has proven highly effective in various applications, such as image recognition, where WRNs have achieved state-of-the-art results. Additionally, WRNs help regularize the training process by allowing better gradient flow through the network, reducing the risk of vanishing or exploding gradients. Given these advantages, WRNs are the technology we will focus on in this thesis to enhance speech quality effectively.

The broad objective of this thesis is to develop new speech enhancement methods using a novel neural network technology and to better understand the internal processes of these networks. This can be divided into the following specific objectives:

- Introduce and evaluate the Wide Residual Network (WRN) architecture specifically for speech enhancement across various conditions
- Enhance the interpretability of neural network models to gain a better understanding of the enhancement process through visualization techniques and exploit this knowledge to achieve additional improvements.
- Explore the insights provided by visualization techniques to develop new strategies that effectively reduce noise and reverberation while maintaining interpretability.

1.3 Thesis organization

This thesis is organized into several chapters, each focusing on different aspects of speech enhancement using neural networks. The structure is designed to guide the reader from a broad introduction to the field, through detailed technical developments, and finally to the conclusions and implications of the research.

- **Chapter 1: Introduction.** Provides an overview of the importance of speech enhancement and sets the context for the research. It discusses the motivation, objectives, and significance of the study.
- **Chapter 2: Overview.** Reviews speech enhancement from two perspectives: the historical development of speech enhancement methods, from early techniques to modern advances, highlighting the evolution of the field, and an algorithmic overview focusing on the main different algorithmic approaches.
- **Chapter 3: Neural Networks for Speech Enhancement.** Explores the use of neural networks in speech enhancement, detailing feature extraction techniques, cost functions, and evaluation metrics used to develop effective models.
- **Chapter 4: Wide Residual Neural Network.** Introduces and evaluates the Wide Residual Network (WRN) architecture for

speech enhancement, demonstrating its effectiveness in various conditions and comparing it to existing methods.

- **Chapter 5: Enhancing Interpretability in Speech Enhancement.** Focuses on improving the interpretability of neural network models, using visualization techniques to gain insights into the enhancement process and evaluating the impact of this understanding on performance.
- **Chapter 6: Progressive Loss Strategies for Enhanced Speech.** Examines the development of progressive loss strategies based on insights from visualization, aiming to reduce noise and reverberation while maintaining interpretability.
- **Chapter 7: Conclusions.** Summarizes the key findings of the research, discusses the implications, and outlines potential directions for future work.

Chapter 2

Overview

The development of speech enhancement techniques has been a crucial area of research, driven by the need to improve the clarity and intelligibility of speech in various noisy environments. The historical context of this field traces back to the mid-20th century, when initial efforts focused on basic noise reduction methods. Over the decades, the field has seen significant evolution, moving from early spectral subtraction techniques to the sophisticated machine learning models that define the current state-of-the-art.

Advances in computing power and the development of new algorithms have been very important in this progression. The integration of statistical methods marked substantial improvements in the first decades of research, but the advent of deep learning models in the 2010s revolutionized the field by providing powerful tools for real-time speech enhancement.

Today, the state-of-the-art in speech enhancement is characterized by the use of advanced methods like convolutional neural networks, generative adversarial networks, and transformers. These techniques have not only improved the quality and intelligibility of speech but also expanded the applications of speech enhancement technology in telecommunications, hearing aids, and voice-controlled systems.

Despite these advances, the field continues to face challenges, such as handling highly variable noise and reverberant environments and developing models that can generalize well across different contexts with

the lowest computational cost. These challenges also present opportunities for future research and innovation.

Modern speech enhancement focuses on developing robust processing methods to improve system performance in real-world scenarios, mainly through larger and more complex systems for better speech signal intelligibility and robustness in recognition systems (Loizou, 2007).

In addition, the field has significantly advanced in understanding how environmental noise and reverberation affect speech signals, leading to the development of more accurate enhancement methods. Techniques now often distinguish between single-channel and multi-channel approaches. While multi-channel methods leverage spatial filtering techniques like beamforming to improve performance in complex acoustic environments(Nakatani et al., 2010), this thesis focuses on single-channel techniques. Single-channel methods are applicable in a broader range of scenarios where only one signal is available, making them more versatile and practical for many real-world applications.

Speech enhancement plays a crucial role across a variety of applications, making it an essential area of research and development. In telecommunications, these techniques are vital for improving voice communication quality over phone lines and VoIP systems, ensuring clear conversations even in adverse noise conditions. This is especially important for mobile communications, where users often encounter noisy environments.

In hearing aids, advanced speech enhancement methods amplify speech while reducing background noise, significantly improving the clarity for individuals with hearing impairments. This enhancement can greatly enhance the quality of life by enabling better communication in daily activities.

Voice-controlled systems, such as virtual assistants and automated customer service interfaces, also benefit immensely from speech enhancement. Clearer speech signals improve the performance of speech recognition systems, making interactions more efficient and natural especially devices located at a certain distance from the user in large rooms where reverberation noise can be harmful. As these systems become

increasingly prevalent, the importance of robust speech enhancement continues to grow.

Additionally, speech enhancement is critical in safety and accessibility technologies. For instance, in emergency response systems, clear communication is essential. Enhanced speech ensures that instructions and information are conveyed accurately in high-noise environments, which can be lifesaving.

Overall, the continuous advance of speech enhancement techniques is driven by their wide-ranging applications, highlighting the importance of this field in both everyday life and specialized contexts.

This chapter is organized in two main sections. The first section provides a detailed timeline of the key developments in speech enhancement, highlighting significant milestones and their impact on the field. The second section describes with more detail methods of methods of speech enhancement, categorized based on their algorithmic relationships into sample generation, mask methods, and spectral reconstruction methods. This structured approach aims to provide a comprehensive overview of the history and current state of speech enhancement techniques.

2.1 Historical Development of Speech Enhancement Methods

2.1.1 Early Beginnings in the 1940s - 1970s

Speech enhancement has been a target of research for several decades, with significant advances in understanding environmental acoustic distortion of speech signals. The journey began in the mid-20th century, driven by the need for improved clarity and intelligibility in noisy environments. This period laid the foundation for many of the techniques and methods that would evolve over the next decades.

One of the earliest significant developments in speech enhancement was the vocoder, created during World War II in the 1940s (Hoffmann, 2010). The vocoder was primarily used to synthesize and encrypt voice communications, ensuring secure transmission of voice signals. This

technology worked by analyzing the speech signal and encoding its essential characteristics, which could then be used to reconstruct the original voice signal at the receiving end. The vocoder represented a significant technological advance and demonstrated the potential of signal processing techniques in enhancing and manipulating speech for practical applications.

In the 1950s, spectral subtraction emerged as one of the pioneering techniques in the field of speech enhancement. Spectral subtraction involves estimating the noise spectrum during non-speech intervals and subtracting it from the noisy speech spectrum. The primary goal of this method is to reduce the impact of background noise on the speech signal, thereby improving its intelligibility and quality. Spectral subtraction establised a basis for more advanced spectral analysis methods. These methods are essential for modern speech enhancement systems (Boll, 1979a).

The initial efforts in speech enhancement during the 1940s and 1950s primarily focused on basic methods to reduce noise and improve intelligibility. These early techniques, such as the vocoder and spectral subtraction, paved the way for the development of more advanced spectral analysis techniques. The evolution from these initial methods marked the beginning of a continuous quest for better and more effective speech enhancement technologies, driven by the growing demand for clear and intelligible speech in various noisy environments.

These foundational techniques were crucial in highlighting the importance of noise reduction and clarity in speech communication, setting the stage for subsequent advances in the field. As research progressed, the understanding of acoustic distortion and the methods to counteract it became more sophisticated, leading to the diverse and advanced speech enhancement technologies we have today.

The 1960s and 1970s marked significant advances in speech enhancement, introducing statistical methods that provided foundational approaches influencing future developments.

Linear Predictive Coding (LPC) was developed in the 1960s. It is a model that LPC models the resonancies of the human vocal tract, with the source representing the vocal cords and the filter representing the

vocal tract. LPC is used for analyzing and creating speech. The filter coefficients obtained by the LPC method capture essential features of the speech and have been used in multiple applications. This technique improved speech compression and synthesis, making speech transmission and storage more efficient (Makhoul, 1975; Atal, 1974; Markel & Gray, 1976).

In the 1970s, Kalman filtering was applied to speech enhancement, offering a recursive solution for estimating the speech signal in the presence of noise. Kalman filters are optimal estimators, assuming that errors have a normal distribution and it operates by predicting the state of a system over time and updating this prediction based on new measurements. When applied to speech enhancement, Kalman filtering continuously estimates the clean speech signal by considering the dynamic nature of both the speech signal and the noise. This method provided a more accurate and adaptive approach to noise reduction compared to earlier techniques, significantly improving the quality of enhanced speech signals (Paliwal & Basu, 1987; Kalman, 1960; Brown, 1983).

The transition from simple noise reduction techniques to complex mathematical models like LPC and Kalman filtering represented significant progress in speech enhancement. These methods allowed for more accurate modeling of speech signals and better handling of noise, setting the stage for future innovations in the field.

2.1.2 Growth in the 1980s - 1990s

During the 1980s and 1990s, statistical methods solidified their place in speech enhancement, with numerous clean speech estimators and spectral filtering techniques being developed. This era marked significant advances in the ability to enhance speech signals in various noise conditions.

The 1980s saw the establishment of spectral filtering techniques as standard approaches in speech enhancement. Techniques such as Spectral Subtraction (Boll, 1979b) and Wiener Filtering (Lim & Oppenheim, 1979) became widely adopted. Spectral Subtraction involves estimating the noise spectrum during non-speech intervals and subtracting it

from the noisy speech spectrum to reduce background noise and improve speech quality. Wiener Filtering, on the other hand, optimizes the trade-off between noise reduction and signal distortion by minimizing the mean square error between the estimated clean signal and the actual noisy signal.

Another significant development in the 1980s was the introduction of Minimum Mean Square Error (MMSE) estimation. This statistical approach aims to minimize the mean square error between the clean and noisy speech signals. The MMSE estimator provided a more sophisticated means of enhancing speech by leveraging the statistical properties of the speech and noise signals. This approach inspired many subsequent methods, such as the Multiplicatively-Modified Log-Spectral Amplitude (MM-LSA) (Malah et al., 1999) and the Optimally-Modified Log-Spectral Amplitude (OM-LSA) (Cohen, 2003) techniques, which further improved speech enhancement performance (Ephraim & Malah, 1984, 1985).

In the 1990s, Hidden Markov Models (HMMs) (Rabiner, 1989) were employed for speech recognition and enhancement. HMMs model the statistical properties of speech signals by considering the temporal variability and sequential nature of speech. These models provided a powerful framework for both recognizing and enhancing speech by capturing the underlying structure and dynamics of the speech signal. HMM-based methods significantly improved the robustness and accuracy of speech enhancement systems, particularly in noisy environments (Veisi & Sameti, 2013).

The integration of statistical methods such as MMSE and HMMs during the 1980s and 1990s significantly improved the ability to enhance speech in varying noise conditions. These methods provided a robust foundation for subsequent developments in speech enhancement, allowing for more accurate modeling and processing of speech signals in diverse acoustic environments.

2.1.3 The 2000s: Machine Learning and Beyond

The 2000s brought significant advances in speech enhancement through learning-based methods, introducing new paradigms that boosted the

field.

Wavelet transform (Daubechies, 1990) emerged as a powerful time-frequency analysis tool in the 2000s, providing better resolution for transient signals compared to traditional Fourier methods. This capability makes it particularly useful for speech enhancement, where the precise localization of signal features in both time and frequency domains is crucial for effective noise reduction and signal reconstruction.

Non-negative Matrix Factorization (NMF) (Mohammadiha et al., 2013; Fan et al., 2014) was explored for its application in decomposing speech signals for enhancement. NMF works by factorizing a matrix into two non-negative matrices, effectively decomposing the speech signal into its constituent parts. This method is particularly advantageous for identifying and separating different sources in a noisy environment, making it a valuable tool for speech enhancement.

The introduction of Deep Neural Networks (DNNs) in the late 2000s had a deep impact in the field by leveraging large datasets and complex architectures to significantly improve performance (Wan et al., 1999). DNNs, with their ability to learn hierarchical representations of data, have been applied to various aspects of speech enhancement, including noise reduction and feature extraction. These models outperform traditional methods by adapting to diverse and complex noise environments, enhancing the overall quality and intelligibility of speech signals.

The application of machine learning, particularly deep learning, transformed speech enhancement, enabling the handling of more complex noise environments and improving the overall quality and intelligibility of speech. The shift from traditional statistical methods to learning-based approaches marked a significant milestone, demonstrating the potential of these technologies to address longstanding challenges in speech enhancement.

2.1.4 The 2010s: Deep Learning and Real-time Processing

Deep learning methods continued to evolve in the 2010s, with new architectures and techniques being developed to further enhance speech signals.

Recurrent Neural Networks (RNNs) were utilized for their ability to model temporal sequences in speech signals (Maas et al., 2012). RNNs are particularly effective in capturing the dependencies in sequential data, making them ideal for tasks such as speech enhancement, where the temporal context of the signal is crucial.

Generative Adversarial Networks (GANs) were applied to generate clean speech signals by learning from noisy examples in an adversarial training setup (Pascual et al., 2017). GANs consist on two main neural networks that are trained simultaneously to create data that is as realistic as possible (Generator) and to distinguish between real data and generated data (Discriminator). This approach has shown significant promise in producing high-quality, clean speech signals from noisy inputs.

Phase-aware methods were developed as an alternative to traditional magnitude-based enhancement techniques (Mowlaee & Kulmer, 2015; Mowlaee et al., 2016), focusing on improving phase information for better speech quality. These methods recognize that accurate phase reconstruction is critical for natural-sounding speech and aim to enhance both the magnitude and phase components of the speech signal.

The refinement and specialization of neural network architectures, such as RNNs, GANs, and phase-aware methods, provided powerful tools for real-time speech enhancement and more natural-sounding results. These advances in deep learning have enabled significant improvements in the ability to handle complex noise environments and enhance the overall quality of speech signals.

2.1.5 The 2020s: State-of-the-Art Techniques

The latest advances in speech enhancement are characterized by the use of advanced neural network architectures and self-supervised learning methods.

Transformers have been employed for their superior capability in modeling long-range dependencies in speech signals (Vaswani et al., 2017; Oostermeijer et al., 2021). Their attention mechanisms allow for capturing complex relationships in the data, making them highly effective for speech enhancement tasks.

Self-supervised learning has leveraged vast amounts of unlabeled data to improve speech enhancement models, making them more robust and versatile. This approach enables models to learn useful representations from the data itself, without relying on extensive labeled datasets (Qiu et al., 2021; Huang et al., 2022). These methods have significantly enhanced the capability to process and enhance speech signals in diverse and complex acoustic environments.

2.2 Algorithmic Overview of Speech Enhancement Methods

In the quest to improve speech intelligibility and quality in noisy environments, a diverse array of speech enhancement techniques has been developed. These techniques can be broadly categorized into three major algorithmic blocks: sample generation methods, mask methods, and spectrum reconstruction methods. This section aims to provide a comprehensive overview of these categories, highlighting the fundamental ideas behind each algorithmic block and illustrating how these concepts are implemented in various speech enhancement methods.

The motivation for this algorithmic overview is to offer a structured understanding of the different approaches to speech enhancement. By grouping the methods into three distinct blocks, we can better grasp the core principles that drive each approach. Sample generation methods focus on creating clean speech samples from noisy inputs, leveraging advanced generative models to achieve high-quality output. Mask methods, on the other hand, involve applying a mask to the noisy speech to isolate and enhance the speech components, a technique that has proven effective in a range of noise conditions. Spectrum reconstruction methods aim to reconstruct the clean speech spectrum from the noisy spectrum, utilizing various statistical and machine learning techniques to achieve this goal.

Understanding these blocks is crucial for understanding modern speech enhancement strategies and identifying each approach's strengths and limitations. This structured perspective not only aids in the theoretical comprehension of the field but also provides practical insights for the development and improvement of speech enhancement technologies.

By analyzing each algorithmic block, we can uncover the innovative techniques that have advanced the field and explore how these methods can be further refined to meet the challenges of real-world applications.

2.2.1 Mask Methods

Mask-based methods work by applying a mask to the noisy speech to isolate the clean speech components. These methods are particularly effective in separating speech from noise in the time-frequency domain, making them widely used in various speech enhancement applications.

Mask-based methods were a key advance of modern speech enhancement, leveraging the concept of masking to separate clean speech components from noisy inputs. These methods operate in the time-frequency domain. By applying a mask to the spectrogram, these methods can selectively attenuate the noise while preserving the speech components, resulting in a cleaner and more intelligible output.

The fundamental principle behind mask methods is the creation of a time-frequency mask that differentiates between speech and noise. This mask can be binary or continuous, depending on the specific method employed. Binary masks, such as the Ideal Binary Mask (IBM) (Wang, 2005), classify each time-frequency bin as either speech or noise, retaining only the bins classified as speech. Continuous masks, such as the complex Ideal Ratio Mask (cIRM) (Williamson et al., 2015), provide a ratio that scales each time-frequency bin according to the proportion of speech and noise present including phase, allowing for more nuanced separation.

The significance of mask methods in today's speech enhancement landscape lies in their effectiveness, versatility and lower computational cost comparing with other methods. These methods are particularly advantageous in scenarios where the noise characteristics are complex and non-stationary, such as in crowded public places, urban environments, or dynamic workspaces.

In real-time communication systems, mask methods enhance the clarity of conversations, ensuring that users can communicate effectively

even in noisy settings while remaining cost-effective. For hearing aids, these methods improve the user's ability to understand speech, significantly enhancing their quality of life. Additionally, mask methods play a crucial role in voice-controlled systems, where accurate speech input is essential for reliable performance.

The continued development of mask methods has been driven by advances in machine learning and deep learning. Modern techniques often involve training neural networks to generate optimal masks based on large datasets of noisy and clean speech pairs. These neural networks learn to identify the complex patterns and structures of speech and noise, resulting in highly effective masks that significantly improve speech enhancement performance. As technology advances, mask methods are expected to become even more sophisticated, leveraging larger datasets and more powerful computational resources.

Time-Frequency Masking (TFM) (Miller, 1947; Yilmaz & Rickard, 2004) utilizes masks derived from neural networks to enhance speech in the time-frequency domain. These masks are generated based on the learned characteristics of clean and noisy speech, allowing for more sophisticated separation of speech components. TFM methods often involve training deep neural networks to predict the ideal mask for each time-frequency bin, resulting in improved speech quality and intelligibility. Two main types of masking have been studied in the literature.

The Ideal Binary Mask (IBM) (Wang, 2005) method classifies each time-frequency bin of the speech signal as either speech or noise. This binary decision results in a mask that retains bins classified as speech and discards those classified as noise. The goal of IBM is to maximize the signal-to-noise ratio (SNR) in the enhanced speech signal by preserving only the components that are most likely to be speech.

The Ideal Ratio Mask (IRM) (Narayanan & Wang, 2013; Ribas et al., 2022) estimates the ratio of the clean speech to the noisy speech for each time-frequency bin. Unlike IBM, which uses a binary decision, IRM provides a continuous mask that scales each bin by a ratio, reflecting the proportion of speech and noise. This method aims to improve speech quality by more accurately representing the contribution of speech and noise in each bin.

2.2.2 Spectrum Reconstrucion Methods

Spectrum reconstruction methods focus on reconstructing the clean speech spectrum from the noisy spectrum. These methods aim to reconstruct the clean speech spectrum from a noisy input, effectively reducing noise while preserving the intelligibility and naturalness of the speech signal. By generating a clean estimated version of the frequency domain, these techniques can target specific noise frequencies, providing a more nuanced and effective enhancement compared to masked approaches which have more problems to reconstruct clean signal when the time-frequency bin is contaminated with noise by masking

The fundamental principle behind spectrum reconstruction methods is to separate the speech components from the noise by estimating and modifying the spectral properties of the signal. This process typically involves transforming the noisy speech signal into the frequency domain using techniques like the Short-Time Fourier Transform (STFT). Once in the frequency domain, various algorithms are applied to estimate the noise and clean speech spectra. The estimated noise spectrum is then subtracted or filtered out, and the remaining clean speech spectrum is transformed back into the time domain to produce the enhanced speech signal. In Appendix A we make a explanation of this process.

These methods are particularly valuable in environments with unpredictable and varying noise conditions, as they provide a flexible and adaptive approach to noise reduction. Furthermore, advances in computational power and machine learning have significantly improved the performance of spectrum reconstruction techniques, making them more effective and efficient than ever before.

Spectral Subtraction (Boll, 1979b; Berouti et al., 1979) is one of the earliest and most straightforward spectrum reconstruction techniques. It works by estimating the noise spectrum during non-speech intervals and subtracting it from the noisy speech spectrum. This method effectively reduces background noise, improving the clarity of the speech signal.

Wiener Filtering (Lim & Oppenheim, 1979) is a more advanced technique that applies a filter designed to minimize the mean square error between the estimated clean signal and the actual noisy signal. This method balances noise reduction and signal distortion, providing a more refined approach to speech enhancement.

MMSE (Ephraim & Malah, 1984; Griffin & Lim, 1984) Estimation aims to minimize the mean square error between the clean and noisy speech signals. Techniques such as MMSE Short-Time Spectral Amplitude (MMSE-STSA) and Log-Spectral Amplitude (MMSE-LSA) (Ephraim & Malah, 1985) estimators fall under this category, offering sophisticated methods for enhancing speech by leveraging statistical properties of the speech and noise signals.

NMF (Mohammadiha et al., 2013; Fan et al., 2014) is used for decomposing the speech spectrogram into non-negative components to separate speech from noise. This method is advantageous for its ability to identify and isolate different sources within the noisy input, making it a powerful tool for speech enhancement.

Deep Neural Networks (DNNs) (Xu et al., 2014) predict the clean speech spectrum from the noisy spectrum using large datasets and complex architectures. These networks are trained to learn the mapping from noisy to clean speech, significantly improving the performance and robustness of speech enhancement systems.

The evolution of spectrum reconstruction methods has been characterized by increasing sophistication and effectiveness in handling noisy speech signals. Early techniques like Spectral Subtraction laid the groundwork, while more advanced methods such as Wiener Filtering and MMSE Estimation provided greater accuracy and adaptability. The introduction of machine learning and deep learning techniques, particularly DNNs, has transformed the field, allowing for more precise and robust speech enhancement. This ongoing evolution reflects the continuous effort to refine these methods to address the challenges posed by complex and dynamic acoustic environments.

2.2.3 Sample Generation Methods

Sample generation methods focus on generating clean speech samples directly from noisy inputs. The key idea behind these methods is to use advanced generative models to produce high-quality speech signals, effectively transforming noisy inputs into clean outputs. These methods are particularly preferred in situations where the noise characteristics are highly variable or unknown, making traditional noise reduction techniques less effective. Applications include real-time communication systems, hearing aids, and any scenario where high-quality speech reconstruction is critical.

Sample generation methods represent a significant advance in the field of speech enhancement, leveraging the power of modern generative models to address the limitations of traditional noise reduction techniques. These methods work by learning the underlying distribution of clean speech signals and using this knowledge to reconstruct clean speech from noisy inputs. This approach contrasts with traditional methods that primarily focus on filtering or subtracting noise, often resulting in artifacts and reduced speech quality.

The process begins with the collection of a large dataset of paired noisy and clean speech samples. Generative models, such as Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs) (Pascual et al., 2017; Bando et al., 2018), and advanced architectures like WaveNet (Qian et al., 2017), are then trained on this dataset. These models learn to map noisy inputs to their clean counterparts, capturing intricate patterns and features of the speech signal that are often lost in noisy environments.

These methods excel in environments where noise characteristics are non stationary, such as public spaces, urban environments, and dynamic work settings. By generating clean speech samples directly, these methods ensure that the enhanced speech is not only intelligible but also natural-sounding, preserving the speaker's original tone and nuances.

As technology continues to advance, the capabilities of generative models are expected to grow, further improving the effectiveness of sample generation methods in speech enhancement. The integration of these

methods into consumer and professional audio devices will likely become standard practice, setting new benchmarks for speech quality in noisy environments.

Variational Autoencoders (VAEs) (Vincent et al., 2010) are another type of generative model used for speech enhancement. VAEs work by encoding noisy speech into a latent space, a lower-dimensional representation that captures the essential features of the speech signal. This latent representation is then decoded to produce clean speech. The VAE framework encourages the latent space to follow a known distribution, allowing for smooth and continuous reconstructions of speech signals. This method leverages the global idea of sample generation by learning a compact representation of speech that is robust to noise, enabling the generation of clean speech from noisy inputs.

Generative Adversarial Networks (GANs) (Pascual et al., 2017; Goodfellow et al., 2014) are a type of neural network architecture that consists of two components: a generator and a discriminator. The generator creates synthetic speech signals from noisy inputs, while the discriminator evaluates the authenticity of the generated speech. The two networks are trained simultaneously in an adversarial process, where the generator aims to produce speech indistinguishable from clean speech, and the discriminator aims to distinguish between real and generated speech. This setup allows GANs to generate high-quality speech signals that effectively mask the noise present in the input.

WaveNet (Qian et al., 2017; Van Den Oord et al., 2016) is a generative model specifically designed for audio signal generation, developed by DeepMind. It models the conditional probability distribution of the audio waveform, allowing it to generate high-fidelity speech signals. WaveNet generates speech one sample at a time, conditioning each new sample on the previous ones, thus capturing the temporal dependencies in the speech signal. This approach makes WaveNet particularly effective in producing natural-sounding speech that closely mimics the characteristics of human speech. By directly modeling the audio signal, WaveNet embodies the sample generation concept, producing clean speech from noisy inputs with high accuracy.

Diffusion-based generative models (Ho et al., 2020; Gonzalez et al., 2024)create new data by gradually transforming random noise into

meaningful patterns through a series of small, guided changes. They share some similarity with the ideas presented in this thesis since they can transform the signal progressively to match a target. This approach can offer a promising new path for research and development in creating clean speech from noisy inputs.

The continuous evolution in this field reflects the growing capabilities of generative models to handle complex noise environments and produce natural-sounding speech, making these methods increasingly relevant in various applications. However, it is important to consider that their computational cost is very high.

Chapter 3

Neural Networks for Speech Enhancement

Neural networks have brought significant advances to the field of speech enhancement, offering sophisticated methods to improve speech clarity and intelligibility in noisy environments. This chapter delves into the critical aspects of data preparation, feature extraction, cost functions, and speech quality measurements necessary for developing effective neural network-based speech enhancement systems.

Data preparation is critical for any machine learning model, particularly in speech enhancement, where the quality and type of data directly influence the model's performance. We will explore the different types of data used in speech enhancement, focusing on the distinction between labeled and unlabeled data. Labeled data is more challenging to obtain but allows for targeted training, while unlabeled data, though easier to acquire, presents its own set of training challenges. Both types require careful preprocessing to maximize their utility, such as classification and cleaning to distill more relevant data.

Next, we will discuss feature extraction, a critical step that involves transforming raw audio data into a format suitable for neural network processing. Various features can be used, such as Mel-Frequency Cepstral Coefficients (MFCC) and Filter Banks, each offering unique benefits for speech enhancement. MFCCs are particularly effective in capturing the phonetic content of speech, while Filter Banks provide a perceptual representation of the spectral properties of the audio signal.

Data augmentation is essential for creating robust and generalizable speech enhancement models. We will examine two main types of noise used for data augmentation: additive noise and convolutive noise. Additive noise involves adding background noise to clean audio to simulate real-world conditions, while convolutive noise, such as reverberation, is simulated using room impulse responses to mimic the effects of acoustic environments. Understanding how to simulate these noise conditions effectively is crucial for training effective models.

Selecting appropriate cost functions is another critical aspect of training neural networks for speech enhancement. Cost functions guide the training process by quantifying the difference between the model's predictions and the clean speech. We will discuss various cost functions, including the widely used minimum mean-square error (MMSE), and explain key concepts such as Short-Time Fourier Transform (STFT) and overlap-add methods, needed by to many speech enhancement algorithms.

Finally, the chapter will address methods for measuring speech quality, an essential part of evaluating the performance of speech enhancement models. We will cover both objective and subjective measures, emphasizing the importance of assessing both intelligibility and naturalness. Objective measures, such as Segmental Signal to Noise Ratio (Segmental SNR), Log Likelihood Ratio (LLR), Speech-to-Reverberation Modulation Energy Ratio (SRMR), Perceptual Evaluation of Speech Quality (PESQ), and Short-Time Objective Intelligibility (STOI), provide quantifiable metrics for evaluation. Subjective measures, often involving listener tests, provide insights into the perceived quality and user satisfaction.

In this chapter, we review the data preparation processes, feature extraction techniques, cost function selection, and speech quality measurement methods necessary for developing and evaluating neural network-based speech enhancement systems. This knowledge will provide to the tools to design experiments and develop models that effectively enhance speech signals in various noisy environments.

3.1 Data Sources and Data Preprocessing

In speech enhancement, the data primarily consists of two types: labeled and unlabeled. Labeled data includes paired noisy and clean speech samples, where the clean speech serves as the target, guiding the model during training to learn the mapping from noisy to clean speech. This type of data is invaluable for precise and targeted training but is difficult to obtain because it requires manual annotation and careful curation. In the context of speech enhancement it would require a very careful and expensive experimental setup with a parallel recording of noise and clean condition, for example a close talk and a room microphone which captures reverberation. Conversely, unlabeled data, which comprises any speech recordings without corresponding clean versions, is significantly easier to collect as it does not necessitate pairing or annotation, leading to an abundance of accessible data given a minimum quality so that we can consider the signals as clean. However the lack of multichannel databases of sufficient size to train large models makes the use of data augmentation over clean dataset the most convenient method. Understanding the importance of preprocessing, and data augmentation is crucial for developing robust speech enhancement models. Understanding these data types and the importance of preprocessing is crucial for developing robust speech enhancement models.

3.1.1 Data Preprocessing

Effective data preprocessing is a critical step in preparing datasets for training neural networks in speech enhancement. This process involves several key tasks, including data classification, data cleaning, and data augmentation, each contributing to the overall quality and robustness of the training data.

3.1.2 Considerations for Data Preparation

When preparing data for training neural networks in speech enhancement, several considerations must be addressed to ensure the effectiveness and efficiency of the training process.

Balance and Diversity

Ensuring a balanced and diverse dataset is essential to prevent model bias and overfitting. A balanced dataset includes a wide range of speech samples with different speakers, accents, speaking styles, and noise conditions. Diversity in the dataset helps the neural network generalize better to unseen data by exposing it to various scenarios during training. Without balance and diversity, the model may become overly specialized in certain conditions, leading to poor performance in real-world applications. In this study, we have utilized reverberation datasets, home noise datasets, and datasets like MUSAN (Snyder et al., 2015), which include continuous and impulsive noises, music, and babble noise.

Scalability

Preparing data pipelines that can handle large datasets efficiently is vital for training robust speech enhancement models. Scalability involves creating automated and streamlined processes for data collection, preprocessing, and augmentation. Efficient data pipelines ensure that large volumes of data can be processed quickly and accurately, enabling extensive and effective training of neural networks. Scalability is particularly important as the size and complexity of datasets continue to grow, necessitating robust infrastructure to manage and utilize this data effectively (Dean et al., 2012).

3.2 Data Augmentation

Data augmentation is essential for improving the dataset with synthetic variations. It helps make the model more robust. This involves creating new training samples by transforming the original audio recordings. In speech enhancement, finding clean-noise signal pairs is difficult. Our goal is to reverse the noise generation process and create realistic clean-noise pairs. We must simulate these pairs as accurately as possible. Key augmentation techniques include:

1. Adding Various Types of Noise: By introducing background noise, such as white noise, pink noise, convolutional noise (Diaz-Guerra

et al., 2021), or real-world environmental sounds, we simulate different noisy conditions that the model will encounter in real-world scenarios. This helps the model learn to distinguish speech from various types of noise.

2. Pitch Shifting: Altering the pitch of the speech without changing its speed can make the model more resilient to variations in speaker characteristics, such as different genders and ages.

3. Time-Stretching: Changing the speed of the audio while preserving the pitch provides additional variations for the model to learn from. This helps in making the model more robust to variations in speaking rates (Ko et al., 2015; Hannun et al., 2014).

These augmentation techniques collectively contribute to creating a diverse and comprehensive training set. They enhance the model's ability to generalize to unseen data, making it more effective in real-world applications.

By systematically applying data classification, cleaning, and augmentation techniques, we can prepare high-quality, diverse datasets that are essential for training robust neural networks for speech enhancement. These preprocessing steps ensure that the data used for training accurately reflects the complexities of real-world conditions, thereby enabling the development of more effective and reliable speech enhancement models. From these techniques discussed in this thesis, we only use the first and the third; however, in other studies, all three have been tested.

3.2.1 Additive Noise

Additive noise refers to any unwanted sound that is directly added to the clean audio signal. This type of noise is typically independent of the speech signal and can include a variety of environmental sounds such as traffic, crowd chatter, white noise, and mechanical noises. The relevance of additive noise in speech enhancement lies in its ability to simulate the diverse and often unpredictable conditions that speech enhancement models will encounter in real-world scenarios. By training models on speech signals with additive noise, we can improve their

robustness and ability to generalize to different noise environments (Virtanen et al., 2012).

To add additive noise to clean audio, we follow these detailed steps:

1. Selecting Noise Types: Choose various types of noise that are representative of real-world conditions the model is expected to encounter. This can include white noise, pink noise, and specific environmental sounds. In the experiments in this thesis we have several noise datasets to choose from.

2. Add the selected noise to the clean audio at different signal-to-noise ratios (SNRs). The SNR is a measure of the level of the desired signal relative to the level of background noise. Adjusting the SNR allows the creation of datasets with varying levels of difficulty. The formulation for adding noise to a clean signal $x(t)$ with a noise signal $n(t)$ is given by:

$$y(t) = x(t) + \alpha n(t) \tag{3.1}$$

where α is a scaling factor that adjusts the noise level relative to the clean signal. The value of α is chosen based on the desired SNR in dBs. For example, to achieve a specific SNR, it can be calculate α using the following relationship:

$$\alpha = \sqrt{\frac{P_x}{P_n \cdot 10^{\frac{SNR}{10}}}} \tag{3.2}$$

where P_x and P_n are the power of the clean signal and noise, respectively. During the training we select SNR randomly in an interval of desired minimum and maximum SNRs.

Additive noise is crucial for creating realistic training scenarios that enhance the model's performance. Here are some common types of additive noise used in data augmentation:

- Environmental Noise: Sounds from specific environments provide context-specific augmentation. For instance, adding traffic noise helps in training models intended for use in urban areas. The type of noise can be stationary, like car, HVAC; and non stationary like street, restaurant, office, etc.

- Babble Noise: Simulates the background noise of multiple people talking simultaneously, which is common in crowded places such as restaurants and conferences. This type of noise is particularly challenging and helps to improve the model's performance in social settings.

By incorporating these types of additive noise into the training data, we can create robust speech enhancement models that perform well under various real-world conditions.

3.2.2 Convolutive Noise

Convolutive noise refers to the modifications of an audio signal caused by reflections and reverberations in an environment. Unlike additive noise, which is directly added to the clean signal, convolutive noise alters the audio signal by convolving it with a Room Impulse Response (RIR). An RIR captures how sound reflects off surfaces such as walls, ceilings, and floors, causing delays and echoes. The impact of convolutive noise on audio signals includes temporal smearing and spectral coloration, making speech less intelligible and more challenging for enhancement models to process effectively. Reverberation, a specific type of convolutive noise, is the persistence of sound in a space after the original sound is produced, caused by multiple reflections from surfaces. It blurs the temporal and spectral features of the speech signal, reducing intelligibility and clarity. Handling reverberation is crucial in speech enhancement because it significantly affects the quality and intelligibility of the enhanced speech (Naylor & Gaubitch, 2010).

To add convolutive noise to clean audio, we follow these steps:

1. Obtaining RIRs: Collect or generate Room Impulse Responses to capture the acoustic characteristics of different environments. RIRs can be recorded in real rooms using specialized equipment, or sourced from existing RIR databases for experimentation. They can also be synthesized using acoustic modeling software. A common method for synthesizing RIRs is the image method (Allen & Berkley, 1979a), and new technologies like GPU-based generation (Diaz-Guerra et al., 2021) offer advanced implementations.

2. Applying RIRs: Convolve the clean audio signal $x(t)$ with an RIR $h(t)$ to produce the reverberated signal $y(t)$. The convolution process is mathematically represented as:

$$y(n) = \sum_{k=0}^{K-1} h(k) \cdot x(n-k) \tag{3.3}$$

Where $x(n)$ is the input signal, $h(n)$ is system impulse response, usually limited to a finite size K by the room simulators, n is the time index and k is a sum variable. This process effectively simulates how the clean speech would sound if it were played in the environment characterized by the RIR.

As mentioned before, the images method is a widely used technique to generate synthetic RIRs by modeling the reflections of sound within a room. This method involves:

1. Modeling the Room Geometry: Define the dimensions and shape of the room, as well as the locations of the sound source and receiver.

2. Simulating Reflections: Calculate the paths of sound waves as they reflect off the room's surfaces. Each reflection path is treated as if it were coming from an "image source" outside the room. These image sources are virtual sources that help simulate how sound waves bounce off surfaces.

3. Generating the RIR: Sum the contributions of all image sources, accounting for the distance traveled and the absorption characteristics of the surfaces. The resulting RIR captures the complex pattern of reflections within the room, which can then be used to add convolutive noise to clean audio signals. As we will explain in Chapter 4 and Chapter 6, in the experiments in this thesis we sample randomly the room size, the positions and other configuration options to increase the variability in the training set.

3.3 Cost functions

A cost function, also known as a loss function, is a mathematical formula used to evaluate how well a neural network's predictions match the actual desired outcomes. In the context of training neural networks for speech enhancement, the cost function plays a critical role in guiding the learning process. By quantifying the difference between the predicted enhanced speech and the actual clean speech, the cost function provides a measure of the model's performance. The goal of training is to minimize this difference, thereby improving the model's ability to produce high-quality enhanced speech.

They provide the feedback necessary for the model to adjust its parameters during training. This adjustment process, typically carried out through optimization algorithms like gradient descent, iteratively reduces the error as measured by the cost function. Without an effective cost function, the model would have no systematic way of improving its performance. For speech enhancement, this means that the chosen cost function directly impacts the clarity and intelligibility of the enhanced speech output (Goodfellow et al., 2016).

The effectiveness of a cost function is crucial for the overall success of the speech enhancement model. Different cost functions can be employed depending on the specific requirements of the task, such as minimizing mean squared error, mean absolute error, or more complex metrics that better capture perceptual aspects of speech quality (Virtanen et al., 2012).

3.3.1 Mean Squared Error (MSE)

Mean Squared Error (MSE) (Bishop, 2006) is a common cost function that measures the average of the squares of the errors between the predicted values and the actual values. It is formulated as:

$$MSE(y, \hat{y}) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{3.4}$$

where y_i is the actual value, $\hat{y}_i$ is the predicted value, and n is the number of observations.

MSE is simple and computationally efficient, making it suitable for a wide range of applications. Its sensitivity to large errors can be beneficial for certain tasks as it encourages the model to focus on reducing larger discrepancies between predictions and actual values.

The primary drawback of MSE is its sensitivity to outliers. Since errors are squared, large deviations disproportionately increase the overall error, which can negatively affect model performance if outliers are present.

In speech enhancement, MSE is often used due to its straightforward calculation and effectiveness in minimizing the overall error between the enhanced and clean speech signals. By focusing on reducing the mean squared error, models can improve the general quality of enhanced speech (Virtanen et al., 2012).

3.3.2 Mean Absolute Error (MAE)

Mean Absolute Error (MAE) measures the average magnitude of errors in a set of predictions, without considering their direction. It is formulated as:

$$MAE(y, \hat{y}) = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{3.5}$$

where y_i is the actual value and $\hat{y}_i$ is the predicted value (Goodfellow et al., 2016).

MAE is robust to outliers because it does not square the errors. This means that large errors do not disproportionately affect the overall error metric, making it more stable when outliers are present.

However, it is less sensitive to smaller errors compared to MSE. This can result in a less precise adjustment of model parameters when fine-tuning to minimize the error.

MAE is particularly useful in speech enhancement when robustness to outliers is critical. It ensures that large errors, such as sudden noise spikes, do not disproportionately affect the training process, leading to a more stable and reliable model performance in noisy environments (Bishop, 2006).

3.3.3 Other Cost Functions

Other notable cost functions include Huber loss and Kullback-Leibler divergence. Huber loss combines the robustness of MAE with the sensitivity of MSE, making it effective in handling outliers and small errors in speech enhancement (Huber, 1992). Kullback-Leibler divergence is useful in probabilistic models for comparing predicted and actual probability distributions, offering an approach to measuring model performance in speech enhancement tasks (Siniscalchi, 2021).

3.4 Feature Extraction

Feature extraction is an important process in speech enhancement, transforming raw audio signals into a format that neural networks can effectively process. This step is essential because it condenses the relevant information in the speech signal, making it easier for models to analyze and enhance the signal. By extracting key features, we can reduce the dimensionality of the data and focus on the most important aspects that contribute to speech clarity and intelligibility. In this thesis these features are used as auxiliary information to the network besides the spectrogram of the noisy signal, this way we intend to provide the network useful information that we expect it can improve the enhacement process.

Various feature extraction methods have been developed to capture different characteristics of speech signals. Among these, Mel-Frequency Cepstral Coefficients (MFCC) and Filter Banks (FB) are widely used due to their inspired design in the human auditory perception. These techniques are integral in many speech processing applications, including automatic speech recognition and speech enhancement.

Beyond MFCC and FB, other methods like Perceptual Linear Prediction (PLP) offer alternative ways to emphasize perceptually important features of speech. PLP, incorporates aspects of human auditory perception to provide a more nuanced representation of speech signals.

In the following subsections, we will explain the feature extraction in more detail.

3.4.1 Filter Banks (FB)

Filter banks are a collection of band-pass filters designed to capture the energy in various frequency bands of the speech signal. Each filter in the bank isolates a specific portion of the frequency spectrum, allowing for detailed analysis of the signal's spectral content.

The process of generating filter bank features involves several steps:

1. Pre-emphasis: This step involves filtering the signal to emphasize higher frequencies. This is done by applying a high-pass filter to the input signal $x(t)$:

$$y(t) = x(t) - \alpha x(t-1) \tag{3.6}$$

   where α is typically set to 0.97.

2. Framing: The continuous speech signal is divided into overlapping frames, typically 20-40 milliseconds long and 10 millisecondss hop, to capture the short-term characteristics of the speech.

3. Windowing: Each frame is windowed using a Hamming window to reduce spectral leakage. The window function $w(t)$ is applied as:

$$y(t) = x(t) \cdot w(t) \tag{3.7}$$

   where,

$$w(t) = 0.54 - 0.46cos(\frac{2\pi t}{N-1}) \tag{3.8}$$

4. Discrete Fourier Transform (DFT): The windowed frames are transformed from the time domain to the frequency domain using the Fast Fourier Transform (FFT) he fast implementation of the DFT:

$$X(f) = \sum_{n=0}^{N-1} x(n) e^{\frac{-j2\pi fn}{N}} \tag{3.9}$$

5. Mel-Filterbank: The power spectrum obtained from the FFT is passed through a Mel-scale filterbank, which consists of triangular filters spaced according to the Mel scale. The Mel frequency

f_{mel} is calculated as:

$$f_{mel} = 2595 log_{10}(1 + \frac{f}{700}) \tag{3.10}$$

6. Log-Amplitude: The log-amplitude of each of the Mel-filterbank outputs is computed, which compresses the dynamic range of the values.

$$E_i = log(\sum_{k=0}^{N-1} |X(k)H_i(k)|^2) \tag{3.11}$$

where E_i is the log energy of the i-th filter, $X(k)$ is the FFT of the windowed signal, and $H_i(k)$ is the frequency response of the i-th filter.

Filter banks are widely used in speech enhancement for their simplicity and effectiveness in capturing spectral information. They provide a perceptual representation of the speech signal's frequency content as we can see in Figure 3.1. By capturing the energy distribution across different frequency bands, filter banks allow speech enhancement algorithms to focus on the most relevant parts of the signal, improving the clarity and intelligibility of the enhanced speech. Their straightforward implementation and computational efficiency make them a popular choice in various speech processing applications.

3.4.2 Mel-Frequency Cepstral Coefficients (MFCC)

Mel-Frequency Cepstral Coefficients (MFCCs) are a set of coefficients that collectively represent the short-term power spectrum of a sound. They are derived from the Mel-frequency cepstrum, which is a representation of the signal's power spectrum on a non-linear Mel scale of frequency (Mermelstein, 1976).

The computation of MFCCs involves the same steps of previous FB features but adding an additional step:

1. Discrete Cosine Transform (DCT): The log-Mel spectrum is transformed into the cepstral domain using the DCT. The resulting

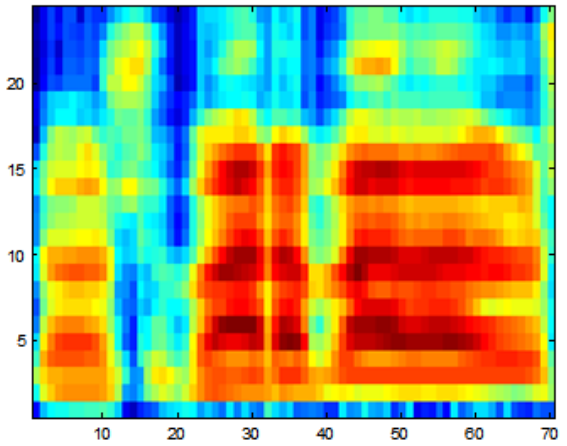


FIGURE 3.1: *Representation of filter bank features*

coefficients are the MFCCs:

$$c_k = \sum_{n=0}^{N-1} log(S_m(n)) cos\left[\frac{\pi k(2n+1)}{2N}\right] \tag{3.12}$$

where $S_m(n)$ is the log-Mel spectrum and c_k are the cepstral coefficients.

MFCCs are widely used in speech technologies due to their ability to mimic the human ear's response. The Mel scale reflects the human ear's perception of sound. This makes MFCCs particularly effective in capturing the perceptual features of speech, improving the performance of speech enhancement systems by aligning more closely with human auditory characteristics (Davis & Mermelstein, 1980). Additionally, MFCCs provide a compact representation of the speech signal, reducing computational complexity while retaining essential information for processing. In Figure 3.2 we can see an example of audio represented by its MFCC features, where we can see that most of the information and energy is concentrated in the lower index cepstrum features.

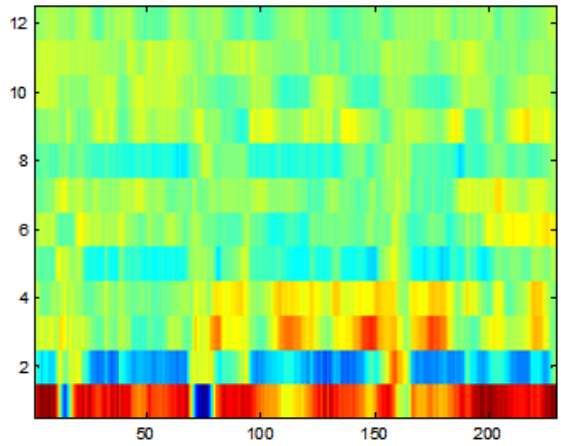


FIGURE 3.2: *Representation of MFCC features*

3.4.3 Other Features

In addition to Mel-Frequency Cepstral Coefficients (MFCC) and Filter Banks (FB), several other feature extraction techniques are employed in speech enhancement. These methods offer unique advantages and are suited to different applications based on their ability to capture specific characteristics of speech signals.

- Perceptual Linear Prediction (PLP): PLP emphasizes perceptually important aspects of the speech signal by modifying the Linear Predictive Coding (LPC) model to better reflect the human auditory system. It was widely used in speech recognition and enhancement for its ability to provide a perceptually accurate representation of speech (Hermansky, 1990).
- Linear Predictive Coding (LPC): LPC represents the spectral envelope of a speech signal in a compressed form using linear predictive models. It is commonly used in speech compression, synthesis, and enhancement due to its efficiency in capturing the speech signal's spectral properties.
- Relative Spectral Transform (RASTA): RASTA emphasizes the temporal dynamics of the speech signal by band-pass filtering the log

spectrum, which enhances robustness to noise and channel distortions. It is particularly effective in noisy and variable acoustic environments, improving the performance of speech recognition and enhancement systems (Hermansky & Morgan, 1994).

3.5 Speech Quality Measurement Methods

Measuring the quality of audio is a critical aspect of speech enhancement, encompassing two primary dimensions: intelligibility and naturalness. Intelligibility refers to how easily the speech can be understood, while naturalness pertains to how natural and pleasant the speech sounds. Although related, these two aspects are not equivalent.

In speech enhancement, accurately measuring audio quality is vital to evaluate and improve the performance of enhancement algorithms. High-quality measurement methods help ensure that enhanced speech is not only intelligible but also sounds natural, which is crucial for applications in telecommunications, hearing aids, and voice-controlled systems. The challenge lies in developing metrics that can objectively quantify these subjective qualities (Hu & Loizou, 2007).

This section will discuss various methods for measuring speech quality, categorized into objective and subjective measures, with further distinctions between those requiring a reference signal and those that do not.

3.5.1 Types of Measures

Evaluating speech enhancement involves two main classifications: objective vs. subjective measures, and reference vs. non-reference measures

Objective vs. Subjective

Objective measures provide quantifiable assessments of speech quality using mathematical models and algorithms. These measures do not rely on human listeners and are typically faster and more consistent. Examples include Mean Squared Error (MSE), Segmental Signal

to Noise Ratio (Segmental SNR), and Short-Time Objective Intelligibility (STOI). Objective measures are particularly useful for large-scale evaluations and when quick, repeatable assessments are needed.

Subjective measures involve human listeners who rate the quality and intelligibility of speech signals. These assessments are often considered the gold standard because they reflect human perception. Methods such as Mean Opinion Score (MOS) and listening tests are common. While subjective measures provide valuable insights, they are time-consuming and may vary between listeners. This thesis did not employ subjective measures due to their higher costs and complexity.

Reference vs. Non-Reference

Reference measures require a clean, original signal to compare with the processed or enhanced signal. They quantify the difference between the two signals to assess the quality of enhancement. Examples include Perceptual Evaluation of Speech Quality (PESQ) and Log Likelihood Ratio (LLR). These measures are effective in controlled environments where the reference signal is available.

Non-reference measures, also known as "no-reference" measures, do not require a clean reference signal. They evaluate the quality of speech based solely on the processed signal. Examples include Speech-to-Reverberation Modulation Energy Ratio (SRMR) and certain no-reference objective measures. These measures are particularly useful in real-world scenarios where a reference signal is not available, for example when the noisy data has been artificially generated, but they are more difficult to obtain in real scenarios.

3.5.2 Segmental Signal to Noise Ratio (Segmental SNR)

Segmental Signal to Noise Ratio (Segmental SNR) measures the signal-to-noise ratio in short, individual segments of the speech signal. It provides a localized assessment of the enhancement process by evaluating the clarity of speech on a segment-by-segment basis. Mathematically, it is defined as:

$$SegmentalSNR = \frac{1}{N}\sum_{i=1}^{N} 10log_{10}\left(\frac{\sum_{n=1}^{M} x_i^2(n)}{\sum_{n=1}^{M}[x_i(n) - \hat{x}_i(n)]^2}\right) \tag{3.13}$$

where $x_i(n)$ is the clean speech segment,$\hat{x}_i(n)$ is the enhanced speech segment, M is the number of the samples in each segment, and N is the number of segments.

Segmental SNR is used to evaluate the clarity of speech enhancement by comparing the power of the clean signal to the power of the noise within each segment. This measure provides a more detailed analysis than the overall SNR by focusing on short-term variations in the speech signal, making it particularly useful for assessing the performance of enhancement algorithms that operate on a frame-by-frame basis.

As an objective measure, Segmental SNR requires a reference signal (clean speech) to calculate the ratio. It helps in quantifying the enhancement performance by providing an average improvement in SNR across all segments. This measure is beneficial in understanding how well the enhancement algorithm performs in different parts of the speech signal, highlighting its effectiveness in both high and low SNR conditions (Hu & Loizou, 2007).

3.5.3 Log Likelihood Ratio (LLR)

og Likelihood Ratio (LLR) measures the difference between the original and enhanced speech signals based on their Linear Predictive Coding (LPC) coefficients. It evaluates how closely the enhanced signal's LPC coefficients match those of the original signal. Mathematically, LLR is defined as:

$$LLR = log\left(\frac{\mathbf{a}_e^T \mathbf{R} \mathbf{a}_e}{\mathbf{a}^T \mathbf{R} \mathbf{a}}\right) \tag{3.14}$$

where $\mathbf{a}$ and $\mathbf{a}_e$ are the LPC coefficient vectors for the original and enhanced signals, respectively, and $\mathbf{R}$ is the autocorrelation matrix of the

original signal. The expression in the numerator calculates the energy of the prediction error when the original signal is modeled with coefficients a, and the denominator when the model coefficientes are a_e. If the enhanced signal spectral envelope is close to the original signal, this ratio will be lower, and in the limit if the enhancement is perfect, the quotient is 1 and the LLR is then 0.

LLR assesses how well the enhanced speech maintains the characteristics of the original signal. By comparing LPC model errors, it provides a measure of the spectral distortion introduced by the enhancement process. A lower LLR indicates that the enhanced signal closely matches the original, preserving the spectral characteristics of the speech.

As an objective measure, LLR requires a reference signal (the original speech) to calculate the ratio. It is particularly useful for quantifying the degree of distortion introduced by speech enhancement algorithms, ensuring that the enhanced signal retains the essential spectral properties of the original speech (Gray & Markel, 1976).

3.5.4 Speech-to-Reverberation Modulation Energy Ratio (SRMR)

Speech-to-Reverberation Modulation Energy Ratio (SRMR) evaluates the amount of reverberation in the speech signal by analyzing the modulation spectrum. It quantifies the ratio of the energy in the speech-modulated components to the energy in the reverberation-modulated components. Mathematically, SRMR is defined as:

$$SRMR = \frac{\sum_{k=1}^{K} \sum_{m \epsilon speech} |M_{k,m}|^2}{\sum_{k=1}^{K} \sum_{m \epsilon reverb} |M_{k,m}|^2} \tag{3.15}$$

where $M_{k,m}$ represents the modulation spectrum coefficients for subband k and modulation frequency m, with speech and reverberation components classified acordingly (Falk et al., 2010b)

SRMR is used to assess the naturalness and clarity of speech in reverberant environments. It helps determine how much the reverberation affects the speech signal, thereby evaluating the effectiveness of dereverberation techniques in enhancing the speech quality.

As an objective measure, SRMR does not require a reference signal. This makes it particularly useful for evaluating real-world recordings where clean references are not available. SRMR provides a robust means of assessing speech quality based on the modulation characteristics of the signal.

3.5.5 Perceptual Evaluation of Speech Quality (PESQ)

Perceptual Evaluation of Speech Quality (PESQ) compares the original and enhanced speech signals using a perceptual model to predict subjective quality scores (Rix et al., 2001; ITU-T Recommendation, 2001). PESQ simulates the human auditory system by evaluating how changes in the speech signal are perceived. Mathematically, PESQ involves transforming the original and enhanced signals into a perceptual domain and computing a difference measure that reflects the perceived quality. The PESQ score is derived from this comparison and typically ranges from -0.5 to 4.5, with higher scores indicating better quality.

The PESQ score is calculated using the following general steps:

1. Perceptual Transform: Transform the original $x(t)$ and enhanced $\hat{x}(t)$ speech signals into the perceptual domain using auditory models.
2. Comparison: Compute the difference between the transformed signals.
3. Aggregation: Aggregate the differences over time and frequency to produce the final PESQ score.

PESQ is widely used in telecommunications to assess the overall speech quality of various codecs and speech enhancement algorithms. It provides a standardized way to measure and compare the performance of different systems based on how listeners perceive the quality of the processed speech. This makes PESQ an invaluable tool for ensuring high-quality speech transmission and enhancement in real-world applications .

As an objective measure, PESQ requires a reference signal (the original speech) to evaluate the enhanced signal. It bridges the gap between objective measurements and subjective listening tests by predicting how users would perceive the quality of the enhanced speech. This predictive capability makes PESQ a trusted metric in both research and industry settings for evaluating speech enhancement performance.

3.5.6 Short-Time Objective Intelligibility (STOI)

Short-Time Objective Intelligibility (STOI) measures the intelligibility of speech by comparing the original and enhanced signals over short time frames (Taal et al., 2011). It quantifies how well the speech enhancement process has preserved or improved the intelligibility of speech in noisy conditions. Mathematically, STOI is computed as the average correlation coefficient between the temporal envelopes of the original and processed speech segments, calculated in short, overlapping time windows.

The STOI score is calculated using the following steps:

1. Short-Time Segmentation: Segment both the original $x(t)$ and enhanced $\hat{x}(t)$ signals into overlapping short-time frames.
2. Temporal Envelope Extraction: Extract the temporal envelopes of each segment.
3. Correlation Calculation: Compute the linear correlation coefficient between the corresponding segments of the original and enhanced signals.
4. Averaging: Average the correlation coefficients over all segments to obtain the final STOI score.

STOI is commonly used to evaluate how well speech enhancement algorithms improve the intelligibility of speech in noisy conditions. It is particularly valuable in scenarios where maintaining or enhancing speech intelligibility is critical, such as in hearing aids, communication systems, and voice-controlled devices. The STOI measure provides a reliable and objective way to assess the effectiveness of different speech enhancement techniques in improving intelligibility under various noise conditions.

As an objective measure, STOI requires a reference signal (the original speech) to compare with the enhanced signal. This comparison allows STOI to quantify the degree of intelligibility improvement provided by the enhancement algorithm, making it a vital tool in the development and evaluation of speech-processing technologies.

Chapter 4

Wide Residual Neural Network

4.1 Introduction

In the preceding chapters, we have traced the development of speech enhancement techniques from their early beginnings to the present day. In this chapter we explore with more detail some models along the line of more modern systems in the overview of Chapter 2, since Deep Neural Networks (DNNs), including Convolutional Neural Networks (CNNs) (LeCun et al., 1998) and Recurrent Neural Networks (RNNs) (He et al., 2016) with Long Short-Term Memory (LSTM) (Hochreiter & Schmidhuber, 1997) units, demonstrated unprecedented capabilities in modeling the complex relationships between data (Maas et al., 2012; Hinton & Salakhutdinov, 2006; Deng & Li, 2013).

Deep learning has provided a dramatic improvement in modern speech enhancement, offering powerful tools to uncover and model the intricate relationships between corrupted and clean speech data. Among these, CNN-based architectures have shown exceptional capability in handling the structural aspects of corrupted speech signals (Fu et al., 2016; Park & Lee, 2017). CNNs excel in capturing spatial hierarchies within the data, making them particularly effective for tasks involving spatially coherent distortions like reverberation.

In addition to CNNs, Recurrent Neural Networks (RNNs) and their advanced variant, Long Short-Term Memory (LSTM) networks, have

proven highly effective in speech enhancement tasks, especially in handling temporal dependencies and dynamic variations in the speech signal (Maas et al., 2012; Weninger et al., 2015; Chen & Wang, 2016; Kinoshita et al., 2017; Gao et al., 2018). These architectures are well-suited for modeling sequential data, allowing them to maintain context over time and effectively mitigate noise and reverberation effects.

Both convolutional and recurrent networks have been further enhanced by incorporating residual connections, which enable the construction of deeper networks with improved convergence and reduced gradient vanishing issues. The residual connection mechanism allows the networks to learn more detailed representations of the speech signal, thereby enhancing the overall performance of speech enhancement systems (Santos & Falk, 2018).

Wide Residual Neural Networks (WRNs) (Zagoruyko & Komodakis, 2017) represent an evolution in deep learning architectures, specifically designed to leverage the benefits of residual connections while increasing the width of the network layers. Unlike traditional deep networks that primarily focus on depth, WRNs expand the layer width, which enhances the network's capacity to learn from data. This architecture is particularly significant in speech enhancement due to its ability to model complex relationships in the spectral domain effectively.

As we show in this chapter, WRNs allow us to address some of the persistent challenges in speech enhancement, such as reverberation and spectral distortion. Reverberation, which occurs when sound reflects off surfaces and causes overlapping echoes, can significantly degrade speech intelligibility. Traditional methods often struggle to accurately separate reverberant components from the desired speech signal. WRNs, with their increased capacity and residual connections, can better capture the nuances of reverberant speech, enabling more effective dereverberation as we show in the experimental section.

This chapter introduces a novel speech enhancement method based on the WRN architecture, utilizing single-dimensional convolutional layers. As we will show, this approach is particularly effective in dealing with reverberation with respect to previous work. By focusing on the log magnitude spectrum, the method makes a direct regression from the reverberant speech spectrum to the clean speech spectrum.

This spectral domain analysis reinforces the importance of low-energy bands, which play a significant role in the perception of speech. This approach ensures that the enhanced speech retains its naturalness and intelligibility, addressing the challenges posed by reverberation and spectral distortion.

The performance of the proposed WRN-based method is evaluated through various speech quality metrics, focusing on both dereverberation levels and the spectral distortion introduced by the enhancement process. These metrics provide a comprehensive assessment of the method's effectiveness in improving speech intelligibility and quality.

To benchmark the proposed method, its performance is compared with the state-of-the-art Weighted Prediction Error (WPE) technique within an experimental framework inspired by the REVERB Challenge. The WPE method, which is based on the LSTM architecture, has demonstrated top performances in handling reverberant speech (Kinoshita et al., 2017).

By comparing the WRN-based approach with the WPE method, this study aims to highlight the advantages of using wide residual networks for speech enhancement. The analysis includes examining how the WRN architecture mitigates the effects of reverberation and spectral distortion, thereby providing clearer and more natural-sounding speech.

This chapter makes significant contributions to the overall thesis by introducing a novel speech enhancement method based on Wide Residual Neural Networks (WRNs). The application of WRNs in this context leverages the benefits of residual connections and increased network width, which together enhance the network's ability to model and improve speech signals corrupted by noise and reverberation. This innovative approach addresses some of the most persistent challenges in speech enhancement, providing a robust framework for improving speech quality.

The chapter also includes a comprehensive performance analysis of the WRN-based method. Various speech quality metrics are used to evaluate the method, with a particular focus on dereverberation levels and the spectral distortion introduced during the enhancement process.

To provide a benchmark, the performance of the proposed method is compared with the state-of-the-art Weighted Prediction Error (WPE) technique, which is based on the LSTM architecture and has demonstrated top performance in handling reverberant speech. This comparison highlights the effectiveness of the WRN approach in mitigating the effects of reverberation and spectral distortion.

The experimental framework for this study is inspired by the REVERB Challenge, ensuring a rigorous and standardized evaluation of the proposed method. This setup allows for a detailed and fair comparison with existing techniques, further underscoring the strengths of the WRN-based approach.

4.2 Wide Residual Network for Speech enhancement

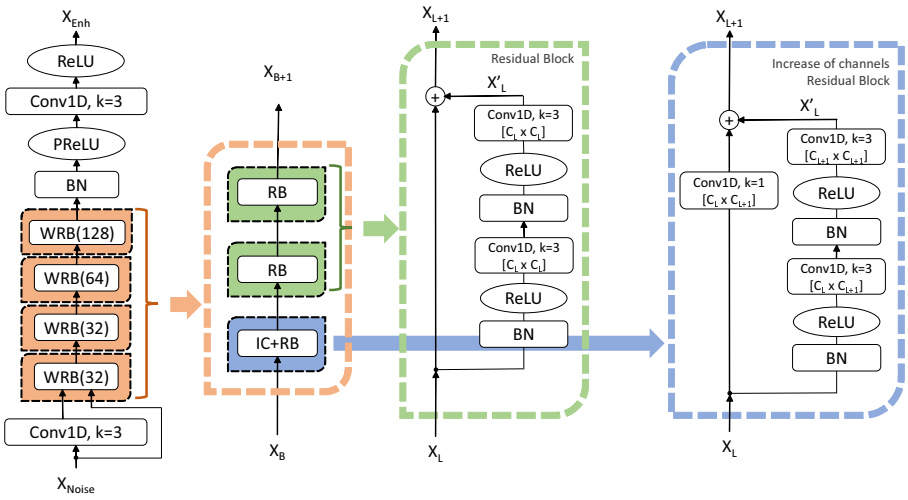


FIGURE 4.1: *Proposed WRN architecture. From left to right, the diagram shows the composition of the network blocks, with C_L representing the number of channels in layer L.*

As we have explained in the introduction the input to the WRN is the log magnitude spectrogram. Then, the network architecture proposed

(Figure 4.1) processes input features with a first convolutional layer followed by four Wide Residual Blocks (WRB). The first WRB processes the output of the first convolutional layer and also its input. Following the WRBs, there is a Batch Normalization (BN) stage and a non-linearity (PReLU: Parametric Rectified Linear Unit). The combination of BN and PReLU blocks provides a smoother representation in regression tasks compared to using ReLU. This WRN architecture increases the number of channels in each stage. In our architecture, we start with 32 channels in the input block and in the first block, but then increase up to 64 channels in the second block and even 128 channels in the last block.

Finally, there is another convolutional layer with a ReLU activation function, reducing the number of channels to 1 and obtaining the predicted enhancement. Each WRB progressively increases the number of channels in its outputs, with the widening operation occurring in the first convolution of the first residual block of each WRB.

To compute the residual connection via a summation operation, the number of channels in both the straight path and the convolutional path must be the same. Therefore, when the number of channels increases, a Conv1D layer with kernel size $k = 1$ is introduced. This acts as a position-wise fully connected layer, aligning the number of channels in the residual path with those in the convolutional path to allow their addition.

In this work, we aim to enhance the logarithmic spectrum of a noisy input signal X_{Noise}. For this purpose, we use the Mean Square Error (MSE) as the training cost function to produce an enhanced signal X_{Enh} that closely resembles the clean reference Y. Drawing from our previous work (Llombart et al., 2018), instead of enhancing each frame individually, we process the entire input signal as a sequence. This approach propagates the accumulated regression error across the entire sentence, rather than frame by frame. This strategy significantly

reduces computational complexity, as each training example is a complete input sequence, rather than hundreds of frames, reusing the context needed by the convolution layers. The cost function, which averages the MSE over all input frames, is described by the equation (4.1):

$$J(Y, X_{Enh}) = \frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{N}\sum_{n=0}^{N-1}\text{MSE}(y_{t,n}, x_{Enh,t,n}) \tag{4.1}$$

where T is the number of frames in the example, N is the feature dimension, $y_{t,n}$ are the frames of Y, and $x_{Enh,t,n}$ are the frames of X_{Enh}.

4.3 Experimental setup

The experimental framework used in this work is based on the REVERB Challenge task[1]. We measure the performance of speech enhancement methods using speech quality metrics. Our goal is to balance dereverberation and avoid adding too much spectral distortion during the enhancement process.

To achieve this, we evaluated the approaches using the official Development and Evaluation sets provided by the REVERB Challenge (Kinoshita et al., 2013). This dataset includes simulated speech, created by convolving clean speech from the WSJCAM0 Corpus (Robinson et al., 1995) with Room Impulse Responses (RIRs) recorded in three different rooms. These rooms have varying reverberation times (RT_{60}) of 0.25, 0.5, and 0.7 seconds, respectively. The recordings were made at two different distances between the speaker and the microphone: a near distance of 0.5 meters and a far distance of 2 meters. Additionally, stationary noise recordings from the same rooms were added to the dataset, maintaining a Signal-to-Noise Ratio (SNR) of 20 dB. This ensures that the simulated data not only has reverberation but also includes realistic background noise.

The dataset also features real-world recordings, captured in a reverberant meeting room with a reverberation time (RT_{60}) of 0.7 seconds. These recordings were made at two distances: near (1 meter) and far (2.5 meters) from the speaker to the microphone, and are part of the

[1] http://reverb2014.dereverberation.com

MC-WSJ-AV corpus (Lincoln et al., 2005). Moreover, we incorporated real speech samples from the VoiceHome dataset, including versions v0.2 (Bertin et al., 2016) and v1.0 (Bertin et al., 2019). The VoiceHome dataset was recorded in actual domestic environments, capturing typical household background noises such as those from a vacuum cleaner, dishwashing activities, or television interviews. This variety ensures the dataset covers a wide range of realistic acoustic conditions.

For training the Deep Neural Network (DNN), we utilized 16 kHz sampled data from multiple sources: Timit (Garofolo et al., 1993), Librispeech (Panayotov et al., 2015), and Tedlium (Rousseau et al., 2014). To enhance the robustness of our model, we augmented this training data. The augmentation process involved adding artificially generated Room Impulse Responses (RIRs) (Allen & Berkley, 1979a) with reverberation times ranging from 0.05 to 0.8 seconds. We also included both stationary and non-stationary noises from the Musan dataset (Snyder et al., 2015), with SNR levels varying between 5 and 25 dB. These noises include different types of background sounds such as music and other speech recordings. Furthermore, we applied time axis scaling at the feature level to create a more diverse training set, thereby improving the generalization capability of our model.

To assess the effectiveness of our proposed WRN speech enhancement method, we compared it with the state-of-the-art dereverberation technique known as Weighted Prediction Error (WPE). WPE is recognized for its effectiveness in reducing reverberation, particularly within the REVERB dataset framework (Kinoshita et al., 2017). For our comparison, we utilized the latest version of the WPE method, which is available online[2] and is also based on Deep Neural Networks (DNN) (Kinoshita et al., 2017). Unlike our approach, WPE employs an architecture centered around Long Short-Term Memory (LSTM) networks. This allows us not only to compare the dereverberation performance but also to assess the differences and advantages of our WRN-based solution from a DNN architecture perspective.

[2] https://github.com/fgnt/nara_wpe

Next, we measured speech quality by evaluating the distortion introduced by the enhancement process. This was done using the Log-likelihood ratio[3] (LLR) (Loizou, 2011). The LLR was computed only for active speech segments, which were identified using a Voice Activity Detection (VAD) algorithm (Ramirez et al., 2004). For this measure, smaller values indicate better speech quality because they show less spectral distortion. Additionally, we assessed the reverberation level of the signal using the Speech-to-Reverberation Modulation Energy Ratio (SRMR) (Falk et al., 2010a). In this case, higher values indicate better speech quality. It is important to note that SRMR can be used with real data, whereas LLR requires both the observed/enhanced signal and a clean reference for computation.

Finally, the front-end of the system begins by segmenting speech signals into frames. These frames are 25 ms, 50 ms, and 75 ms long, with each frame being windowed using a Hamming window. The frames are created every 10 ms. This approach helps to capture as much of the reverberant impulse response as possible within each window, without losing the temporal resolution of the acoustic events. For each frame segment, we compute and stack three types of acoustic feature vectors to form a single input feature vector for the network. These feature vectors include a 512-dimensional Fast Fourier Transform (FFT), log magnitude Mel filterbanks of dimensions 32, 50, and 100, and cepstral features that match the dimension of the corresponding Mel filterbank. This combination ensures that a wide range of spectral and temporal information is captured.

After computing the feature vectors, we normalize each vector by its variance. This step helps in stabilizing the training process by ensuring that all features are on a similar scale. The input features are generated and augmented dynamically, processing continuous blocks of 200 samples. This allows for efficient time-domain convolutions. The network architecture includes four Wide Residual Network (WRN) blocks, each with a widening factor of 8. The AdamW algorithm is used to train the network, which combines the benefits of Adam optimization with weight decay regularization. We also use Parametric Rectified Linear Units (PReLUs) as the activation function, as described by (He et al.,

[3]Originally known as Itakura distance

2015). PReLUs help in addressing the dying ReLU problem by allowing a small, trainable gradient when the unit is not active.

4.4 Analysis of Results and Insights

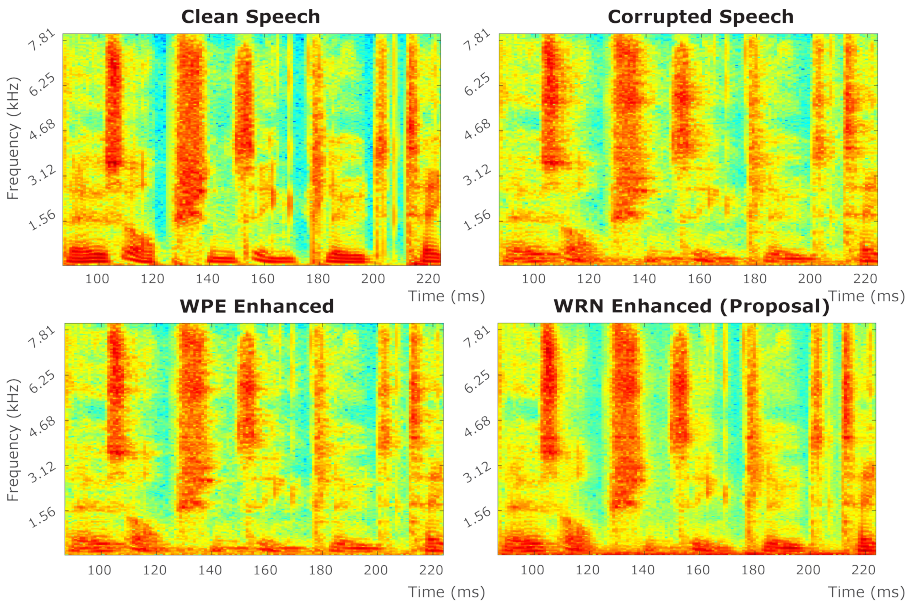


FIGURE 4.2: *Visual example of enhancement applied to a signal from the REVERB Dev dataset.*

Figure 4.2 provides a qualitative example of the enhancement performance using the signal c31c0204.wav from the REVERB Development dataset. This signal has a reverberation time (RT_{60}) of 0.25 seconds and a speaker-to-microphone distance of 200 cm.

In the top-right side of the figure, we can see the spectrogram of the corrupted speech. The distortion caused by reverberation is evident. Reverberation leads to a significant temporal spreading of the power spectrum during active speech segments. This spread blurs the speech signal, making it less clear and harder to understand.

The bottom left part of the figure shows the enhanced speech using the Weighted Prediction Error (WPE) method. WPE reduces some of the

reverberation effects, but it does not completely eliminate them. The enhanced speech is clearer than the corrupted version, but there are still noticeable artifacts.

The bottom right part of the figure displays the enhanced speech using our WRN method. The WRN method more effectively reduces the reverberation and reconstructs the speech signal with greater accuracy as we will show later with the objective metrics evaluated. Compared to WPE, the WRN method produces a cleaner and more natural-sounding speech signal. This demonstrates the superior performance of the WRN approach in handling reverberation and improving speech quality.

4.4.1 Spectral Distortion Analysis

Table 5.1 shows the speech quality results in terms of distortion measured by the Log-Likelihood Ratio (LLR) distance for simulated speech samples. The first row represents the unprocessed reverberant speech. This row serves as a baseline to compare the quality of the enhanced speech signals produced by the WPE method and our proposed WRN enhancement method.

Both WPE and WRN methods are based on Deep Neural Networks (DNNs) and are designed to enhance corrupted speech data. However, our WRN method significantly reduces spectral distortion more effectively than WPE.

The LLR values in the table illustrate this improvement. Lower LLR values indicate better speech quality due to reduced distortion. For the REVERB Development (REV-Dev) and Evaluation (REV-Eval) datasets, the unprocessed speech has LLR distances of 0.63 and 0.64, respectively. The WPE method improves these values to 0.60 for both datasets. In contrast, our WRN method achieves even lower LLR distances of 0.50 for REV-Dev and 0.51 for REV-Eval, highlighting its superior performance in enhancing speech quality.

TABLE 4.1: *LLR distance in simulated reverberated speech samples from REVERB Dev & Eval datasets.*

Methods	REV-Dev	REV-Eval
Unprocessed	0.63	0.64
WPE (Drude et al., 2018)	0.60	0.60
WRN	0.50	0.51

4.4.2 Robustness of WRN in Simulated and Real Environments

Table 4.2 presents the average Speech-to-Reverberation Modulation Energy Ratio (SRMR) results for both simulated and real speech samples under various conditions. The first column shows the SRMR values for the unprocessed speech data, serving as a baseline for comparison. The cells highlighted in gray indicate the best results for each dataset.

TABLE 4.2: *Speech quality through SRMR results for simulated and real reverberated speech samples.*

Datasets	Unprocessed	WPE (Drude et al., 2018)	WRN
Simulated			
REVERB Dev	3.67	3.90	4.75
REVERB Eval	3.68	3.91	4.63
Real			
REVERB Dev	3.79	4.17	4.79
REVERB Eval	3.18	3.48	4.20
VoiceHome v0.2	3.19	3.28	5.03
VoiceHome v1.0	4.51	4.96	5.92

The WRN method outperforms the baseline methods across all evaluated datasets. This consistent performance across different datasets demonstrates the robustness of the WRN method. Unlike some methods that may be fine-tuned for specific datasets, the WRN model seems to better generalize, which is a desirable trait for speech enhancement.

These positive results suggest that the WRN method is not only effective in simulated environments but also in real-world scenarios. It is noteworthy that the WRN model, trained with artificially synthesized reverberation, also excels in handling real reverberated speech. This

indicates the potential for practical applications where real-world reverberation conditions are present.

Reverberation Time and Room Size Effects on Speech Quality

Figure 4.3 shows how the SRMR results change with increasing levels of reverberation for different room sizes: $Room1$ with $RT_{60} = 0.25s$, $Room2$ with $RT_{60} = 0.5s$, and $Room3$ with $RT_{60} = 0.75s$.

The proposed WRN method consistently achieves higher speech quality than the reference methods across all conditions. The results demonstrate that the WRN method is robust, as it shows less variability in SRMR values across different reverberation times (RT_{60}).

Additionally, the improvement in speech quality using the WRN method becomes more pronounced as the reverberation time increases. This means that the method is particularly effective in more challenging reverberation conditions. However, in $Room1$ with $RT_{60} = 0.25s$, there is less room for improvement, making it harder to enhance the speech quality significantly in this scenario.

Performance in Near-Field and Far-Field Conditions

Figure 4.4 shows the average SRMR results for both far-field (250 cm) and near-field (50 cm) conditions in the simulated REVERB Development and Evaluation datasets.

In these tests, the WRN method significantly outperformed the WPE baseline. For far-field conditions, WRN achieved a 34.88% improvement over WPE. In near-field conditions, WRN showed an 8.44% improvement. These results highlight that the WRN method is particularly effective in far-field scenarios, which are generally more challenging due to the increased distance between the speaker and the microphone.

The performance of the WRN method in far-field conditions demonstrates its robustness and capability to handle difficult reverberation scenarios effectively.

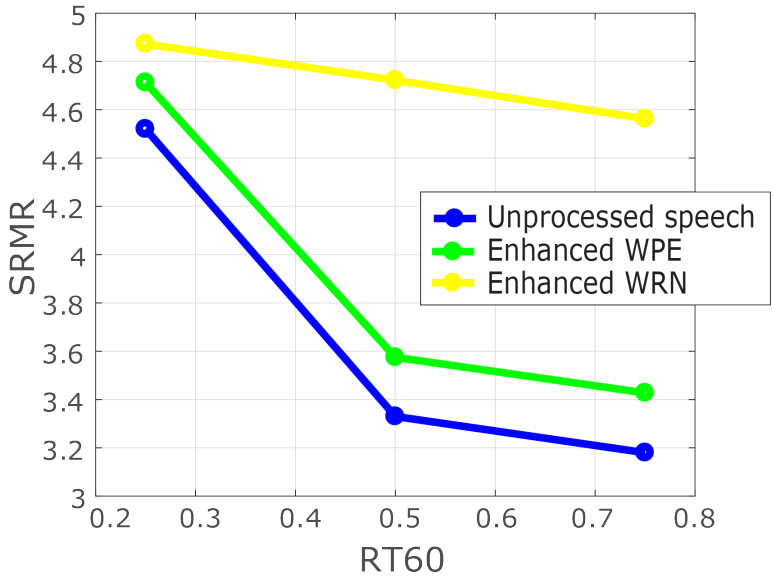


FIGURE 4.3: *Speech quality through SRMR measure for different reverberation levels in simulated reverberated speech samples from REVERB Dev & Eval datasets.*

4.4.3 Challenges of Training-Testing Misalignment

As we observed earlier, enhancing speech quality in *Room*1 with $RT_{60} = 0.25$ and a near speaker-microphone distance was particularly challenging. These conditions involve low reverberation, which provides limited room for improvement. Because of this, focusing data augmentation on these specific conditions during network training could potentially enhance performance.

One of the challenges faced was the lack of precise room size values in the test dataset description. The WRN training data included estimated small room sizes based on reasonable assumptions. However, these estimates might not have been accurate enough for the actual small size of *Room*1. This mismatch likely contributed to the difficulties in achieving better enhancement in these scenarios.

Moreover, the training data augmentation configuration assumed that the speaker and microphone could be randomly placed throughout the

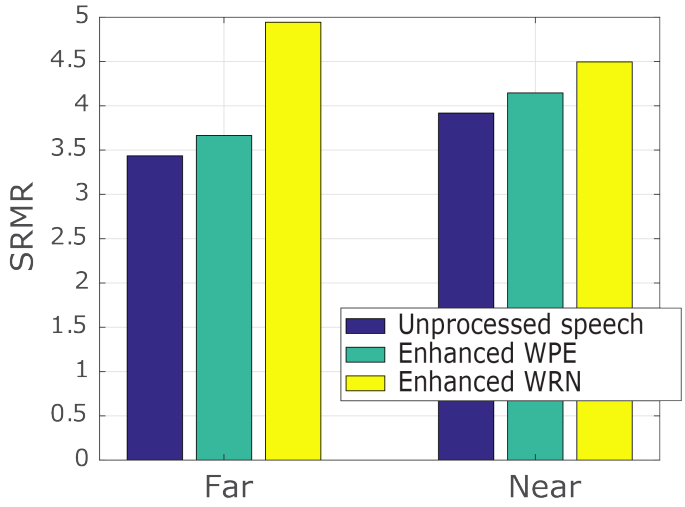


FIGURE 4.4: *SRMR results for simulated reverberated speech in near- and far-field conditions from REVERB Dev & Eval datasets.*

room. This setup modeled the environment with a uniform data distribution. Consequently, this approach resulted in a low probability for encountering specific test distances such as near (50 cm) and far (250 cm).

To address these issues and improve performance in these challenging scenarios, future training data should incorporate smaller room sizes and, in general, a wider set of room configurations in the data augmentation. Additionally, the function used to model speaker-microphone distances should be adjusted. Increasing the probability of encountering specific distances during training can help the model better handle these conditions. However, it is crucial to avoid overfitting the training data to specific scenarios and our design policy during the experiments has been to have a balanced approach, ensuring both generalization and alignment with test data.

4.4.4 Advancements Over Existing Methods

The experimental results clearly demonstrated that the proposed WRN architecture outperformed the reference WPE method in various conditions. While the RNN-LSTM architecture used in WPE is powerful and has its advantages, our WRN approach, which combines CNNs with residual connections, offered more expressive representations of reverberant speech.

The structure of the WRN method allows it to enhance the entire utterance by applying convolutions across the full temporal domain of the signal. This is particularly beneficial as the depth of the network increases, allowing for more complex and detailed feature extraction. In contrast, the WPE method, based on RNN-LSTM, only considers the previous context when processing the speech signal. This limitation can affect the enhancement quality, as it does not account for future context, which can be crucial for certain speech characteristics.

Our WRN method, on the other hand, implements this forward-looking perspective through convolutional layers, which consider all the context around the analysis window. This approach results in a more comprehensive enhancement, improving the overall speech quality. The WRN architecture successfully reconstructed the clean speech signal, achieving higher speech quality than the WPE method. It maintained a proper balance between the level of dereverberation and the amount of spectral distortion.

Furthermore, these positive results were not limited to simulated environments. They were also validated through tests on real distorted speech, demonstrating the model's strong generalization capability. The ability to perform well on real-world data highlights the practical applicability of the WRN method, making it a robust choice for various speech enhancement tasks.

4.5 Conclusions

This chapter introduced a novel speech enhancement method based on a Wide Residual Network (WRN) architecture. This method leverages

the powerful representations provided by a wide topology of Convolutional Neural Networks (CNNs) with residual connections. The results demonstrated that the WRN method outperforms the state-of-the-art RNN-LSTM-based method, known as Weighted Prediction Error (WPE), especially in far-field reverberated speech across three different room sizes.

The residual connections were particularly beneficial because they allow the network to maintain a linear shortcut for the signal, while the non-linear path can enhance the signal by adding or subtracting corrections at specific steps. This characteristic is highly valuable in practical applications, where the system might encounter a variety of challenging conditions (Ribas et al., 2016).

Although the results are promising, there is room for improvement. The subsequent chapters of this thesis address these improvements and further explore the potential of the WRN method. In particular, Chapter 5 explores how we can visualize the process of enhancing speech. This visualization aims to provide deeper insights into how the WRN model processes and enhances speech signals. By visualizing different stages of the enhancement process, we can better understand the transformations occurring within the network, addressing the "black box" nature of deep learning models.

The visualization approach presented in Chapter 5 allows us to track the enhancement process step-by-step through visualization probes at each network block. This method helps us supervise the enhancement process and gather relevant details on how it is performed. Such insights are crucial for identifying which steps are most meaningful in the enhancement process and which can be optimized or discarded. This contributes to achieving a proper trade-off between accuracy and computational effort.

In this chapter, we have shown some encouraging initial results of the WRN architecture. The next chapters will provide a detailed analysis and visualization of the enhancements. These chapters introduce methodological improvements that enhance the model's performance and interpretability. These advancements collectively ensure that the WRN method remains robust and effective for various speech enhancement applications.

Chapter 5

Enhancing Interpretability in Speech Enhancement through Deep Learning Architectures

5.1 Introduction

Fields of speech processing like speech enhancement (SE) have been dramatically improved with the advent of Deep Neural Networks (DNN). These techniques have become the center of interest of the research community for their ability to extract and process information, showing significant improvements over traditional methods (Xia & Bao, 2013; Feng et al., 2014; Tu & Zhang, 2017; Karjol et al., 2018).

Previous chapters have discussed various methodologies and the evolution of neural networks that aid in the enhancement of speech. However, despite their success, a significant challenge persists—the opacity of these methods. Commonly known as the "black box" problem, this lack of transparency in DNN operations makes it hard to understand how they work. As a result, people often have to rely on empirical methods.

This chapter aims to address these concerns by focusing into the interpretability of neural networks, an emerging field that seeks to unravel how DNNs manage feature selection and decision-making processes.

The increasing focus on interpretability is evidenced by its prominence in recent top scientific conferences.

Motivated by the need for clarity and precision in SE solutions, this chapter introduces an innovative SE architecture that employs a feature-mapping strategy. This architecture allows for the visualization of the enhancement process at each step of the network. Such visualization not only aids in understanding the transformations occurring within the network but also enables the evaluation and refinement of the process, even post-training. This adaptability ensures an optimal balance between accuracy and computational efficiency tailored to specific application needs.

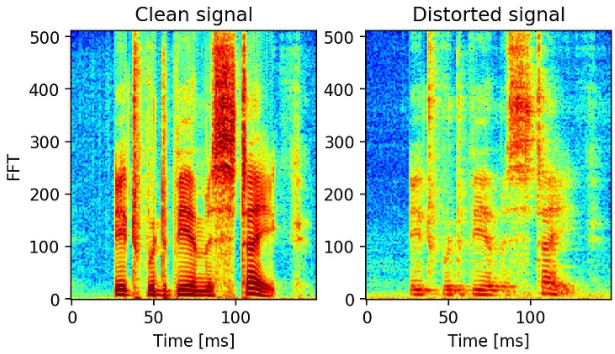
Building on the robust foundations of Residual Networks (RN) discussed in the previous chapter and leveraging one-dimensional convolution layers, the proposed architecture aims to match state-of-the-art performance while providing a granular view of the network's step-by-step processing. The utility of residual connections is explored in depth, highlighting their dual role in linear and non-linear path enhancements, which dynamically adjusts to the distortion levels in the signal (Qian et al., 2017; He et al., 2016; Park & Lee, 2017).

This chapter extends the discussion on deep learning in speech enhancement from previous chapters by providing a detailed exploration of a novel RN-based architecture. It also offers a comprehensive analysis of its performance on reverberated speech through various speech quality measures.

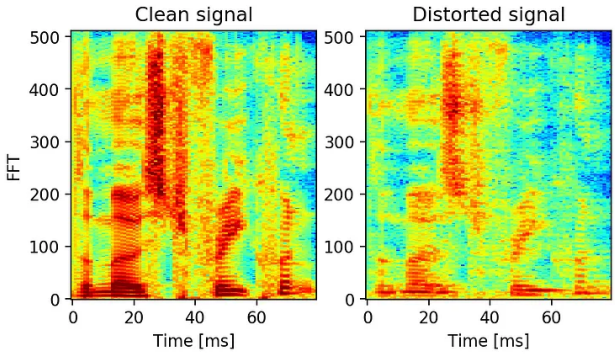
5.2 Proposed architectures

This section introduces the architectures designed for enhancing speech signals, particularly focusing on addressing the challenges depicted in the following figure. Figure 5.1 visually demonstrates the initial state of audio signals before processing. On the left the reference clean signal and on the right the distorted signal with reverberation. These visualizations serve as a clear example of the types of noise and distortions that our proposed architectures aim to mitigate. By analyzing these differences in time-frequency patterns displayed in the log magnitude spectrogram, we can better adjust our methods to enhance the

audio. We will go into more detail about how we do this for each architecture later in this chapter.



(A) Sample 1



(B) Sample 2

FIGURE 5.1: *Comparative log magnitude spectrogram analysis of two audio samples, illustrating the clean and corresponding noisy signals. These visualizations highlight the challenges faced in speech enhancement tasks discussed later in this chapter.*

5.2.1 Constant Channel Residual Network

In our effort to improve speech without losing detail in each processing step, we have developed a Residual Network (RN) that keeps the same number of channels through all residual connections. We call this design the Constant Channel Residual Network (*CCRN*).

Figure 5.2 displays our system. It uses various input sources to capture a wide range of signal representations. We aim to keep as much of the reverberant characteristics as possible. Using a window that is too short compared to the effective RIR length of the channel would create unwanted artifacts. Therefore, we provide time-frequency analysis with several window lengths as input to the system.

The initial processing of speech signals splits them into segments using 25, 50, and 75 ms Hamming window frames, repeating every 10 ms. Each segment calculates three types of acoustic features based on the window size. These features include the the log magnitude of the 512 - dimensional Fast Fourier Transform (FFT) and Mel filterbank & cepstral features of dimensions 32, 50, and 100. These are then combined into a single input feature vector for the network, totaling 876 dimensions. Each feature vector undergoes variance normalization for consistency.

The network's first layer processes the input features and is followed by 14 Residual Blocks (*RB*). This layer matches the input dimension with the number of input channels and uses the dimension of the logarithmic spectrum for the output channels. Each *RB* includes a Batch Normalization (BN) layer (Ioffe & Szegedy, 2015), a Parametric Rectified Linear Unit (PReLU) (He et al., 2015), and a 1-dimensional convolution layer with a kernel size of 3. The number of output channels remains consistent with the input, ensuring stability in feature representation. The combination of *BN* and *PReLU* yields a smoother output ideal for regression tasks, superior to the typical *ReLU*.

The residual connection simply adds the input of the *RB* to its output, which helps in maintaining the integrity of the signal through each block. Our objective is to derive the clean signal's logarithmic spectrum from the noisy input, treated as a continuous sequence rather than segmented frames, based on insights from prior work (Llombart et al., 2018).

We apply a Mean Square Error (*MSE*) loss function of the predicted output with respect to the clean signal across frames as follows:

$$J(Y, X_{S,L}) = \frac{1}{N} \sum_{n=0}^{N-1} \frac{1}{T} \sum_{t=0}^{T-1} MSE(y_{n,t}, x_{S,L,n,t}) \tag{5.1}$$

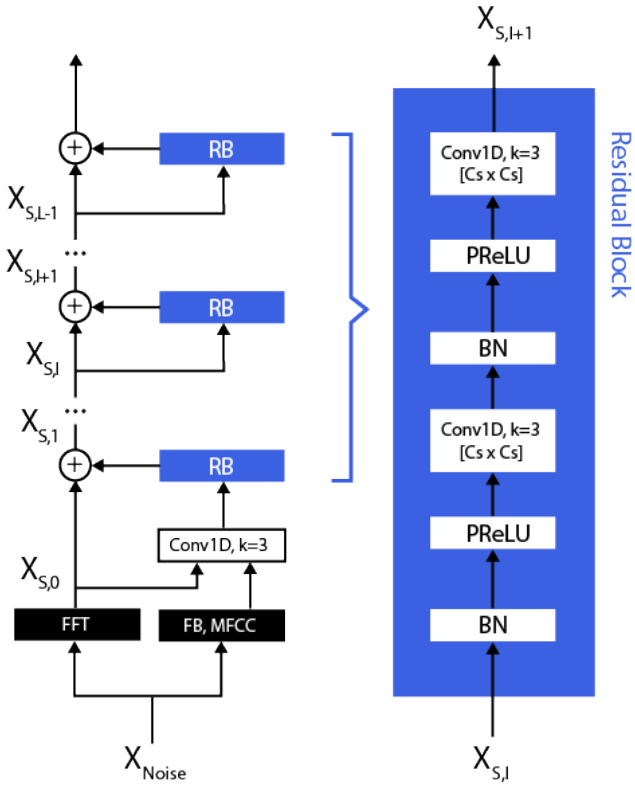


FIGURE 5.2: *Constant Channel Residual Network* (CCRN) *architecture for progressive speech enhancement.* $L = 14, C_S = 512$

where N represents the feature dimension, T is the sequence length, $y_{n,t}$ are the frames of Y the clean signal log magnitude spectrogram, and $x_{S,L,n,t}$ are the frames of $X_{S,L}$, the last block L output.

To observe the enhancement process in action, we position an output probe at each block, made possible by keeping the channel count consistent across all *RB*s. Figures 5.3 and 5.4 illustrate how the spectrum evolves at various stages of processing.

Interestingly, a standard convolution layer mixes all input channels, affecting each output channel differently. Consequently, the 512 - dimensional output matrix does not accurately represent a true spectrum. Certain frequency channels gather a lot of the spectral information,

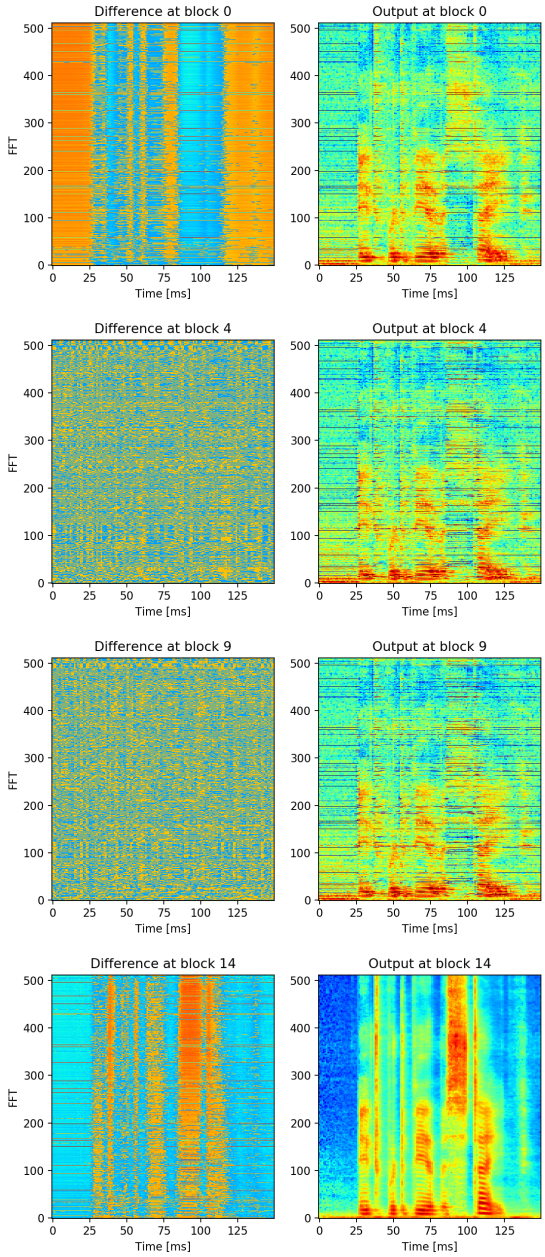


FIGURE 5.3: *Speech enhancement reconstructed output of progressive steps with* CCRN *in a signal sample 1.*

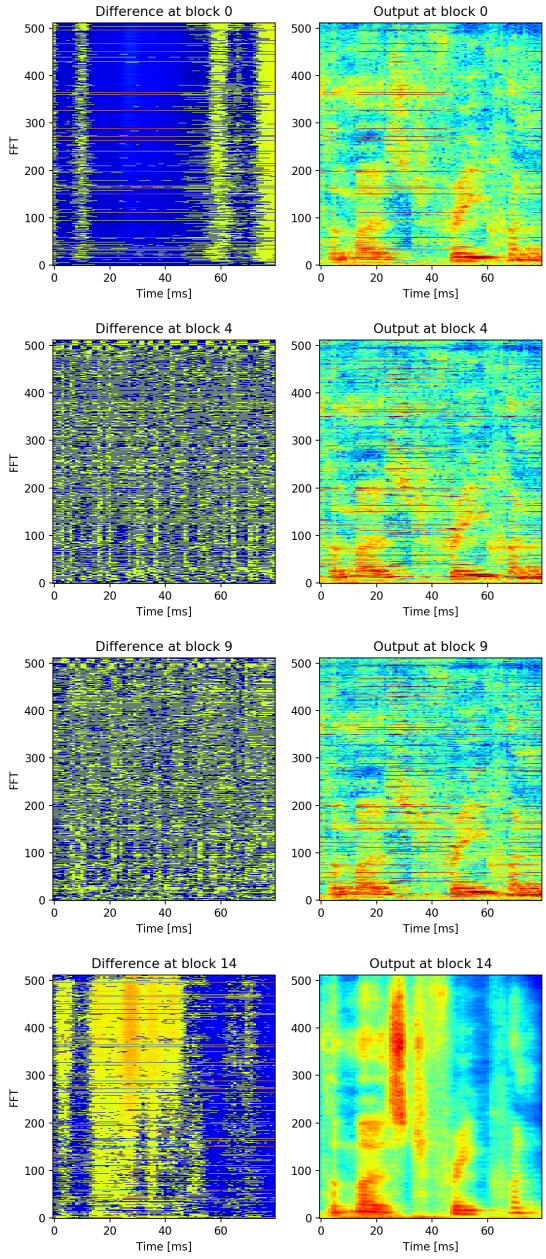


FIGURE 5.4: *Speech enhancement reconstructed output of progressive steps with* CCRN *in a signal sample 2.*

making them appear more defined, while others look blurred. This selective emphasis, noted in (Santos & Falk, 2018), indicates that the network prioritizes specific frequencies based on the varying levels of distortion across the spectrum.

At the beginning and end of the processing chain, notable changes are evident. Initially, the network modifies the spectrum by either intensifying or blurring channels depending on the outputs from the convolution. This transformation is focused primarily on sections containing speech, as shown by the significant alterations in these areas compared to the relatively unchanged non-speech sections. Throughout the depth, as the processing progresses through the intermediate blocks, this clear distinction between speech and non-speech segments becomes obscured due to the scrambling of network channels, making structural patterns difficult to discern. Remarkably, by the final step, the network reorganizes and enhances the signal, effectively reinstating the separation between speech and non-speech sections, suggesting a reordering of channel organization forced by the reconstruction loss objective.

This detailed observation of the network's processing at various stages underscores the dynamic nature of speech enhancement within our proposed architecture, highlighting its ability to adapt and refine the audio signal thanks to several additive correction steps throughout the residual blocks.

5.2.2 Constant Channel Residual Network with State Path

The *CCRN* architecture aims to improve the input signal while preserving its log-spectral integrity progressively. However, despite incorporating a shortcut that allows the input to pass through with minimal changes, the training of the *CCRN* sometimes results in a disordered spectral representation. As discussed in section 5.2.1, it appears that much of the information is concentrated in certain channels.

To enable the input to travel along the residual path without altering its representation, we have introduced a state path between the Residual Blocks (*RBs*). This design allows the signal representation generated by the network to have its own distinct pathway.

Additionally, this state path facilitates an increase in the number of channels at each layer while keeping the number of channels in the residual path consistent. We refer to the updated design illustrated in figure 5.5 as the Constant Channel Residual Network with State path (*CCRN-State*).

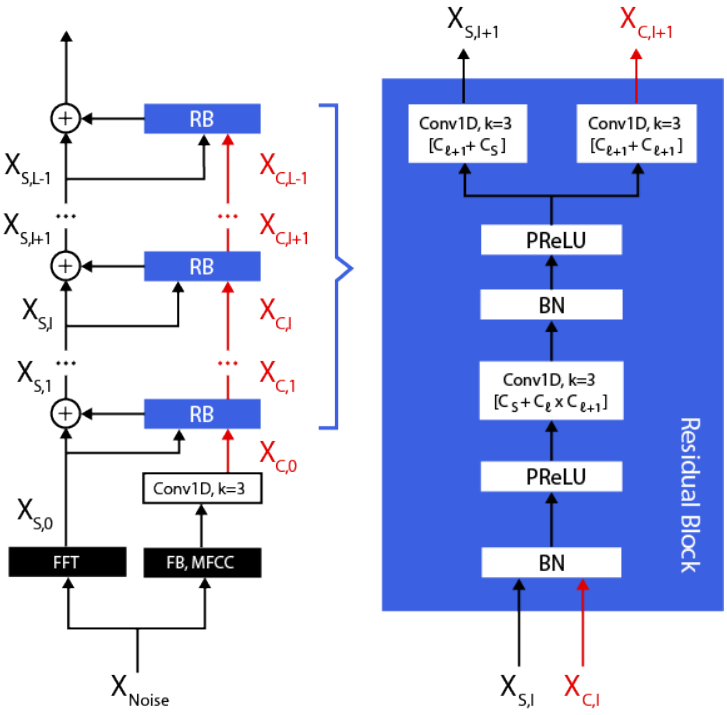


FIGURE 5.5: *Constant Channel Residual Wide Network with State path* (CCRN-State) *architecture for progressive speech enhancement.* $L = 14, C_S = 512, C_l = 32 * l, l \in [1, L]$

In this architecture, the channels from both paths are combined at the input of each block. Drawing inspiration from Wide Residual Networks (Zagoruyko & Komodakis, 2016), we increase the number of channels in the first convolution of each block. Then, to separate the residual and state paths, we employ two convolutional layers at the output of each block. One layer reduces the number of channels back to the dimension of the residual connection, maintaining the original behavior of the architecture. The other layer extends the state path for

use in the subsequent block.

Despite these modifications, as we can see in Figures 5.6 and 5.7 qualitative results once again reveal a spectrum with disorganized frequency channels, reflecting a pattern similar to that observed in the previous architecture.

5.2.3 Progressive Supervision

To ensure each network step accurately reconstructs the signal, we incorporate a Mean Square Error (*MSE*) cost term at every block output. This approach, inspired by similar strategies in classification tasks (Lee et al., 2015), adapts well to our regression task. The idea of progressive fitting and additive reconstruction can also be found in modern boosted decision tree algorithms (Chen & Guestrin, 2016).

In equation (5.2), we augment the training cost by adding the *MSE* between the clean reference and each block output $X_{S,l}$ for all $l \in [1, L]$. This addition to the cost function is controlled by a weighting factor α. In our experiments, we determined $\alpha = 0.1$ based on development trials. We refer to this method as *Progressive Supervision*, as it ensures the network progressively refines the signal at each stage.

$$J_{PS}(Y, X_{S,L}) = J(Y, X_{S,L}) + \alpha \frac{1}{L} \sum_{l=1}^{L} J(Y, X_{S,l}) \tag{5.2}$$

Figures 5.8 and 5.9 displays two spectrogram examples demonstrating the evolutionary enhancement pattern achieved through this method.

The initial blocks focus on more noticeably distorted parts of the spectrum, such as reverberation trails. This suggests the network learns to distinguish between distorted and normal speech patterns. Notably, the network prioritizes the valleys of the spectrum, gradually eliminating finer distortions.

The process gradually reduces the reverberation effects, making the sound clearer and smoother. This helps avoid sharp changes in the sound that might cause unwanted noises, like static. By looking at the differences between the input and the output in each block in Figures

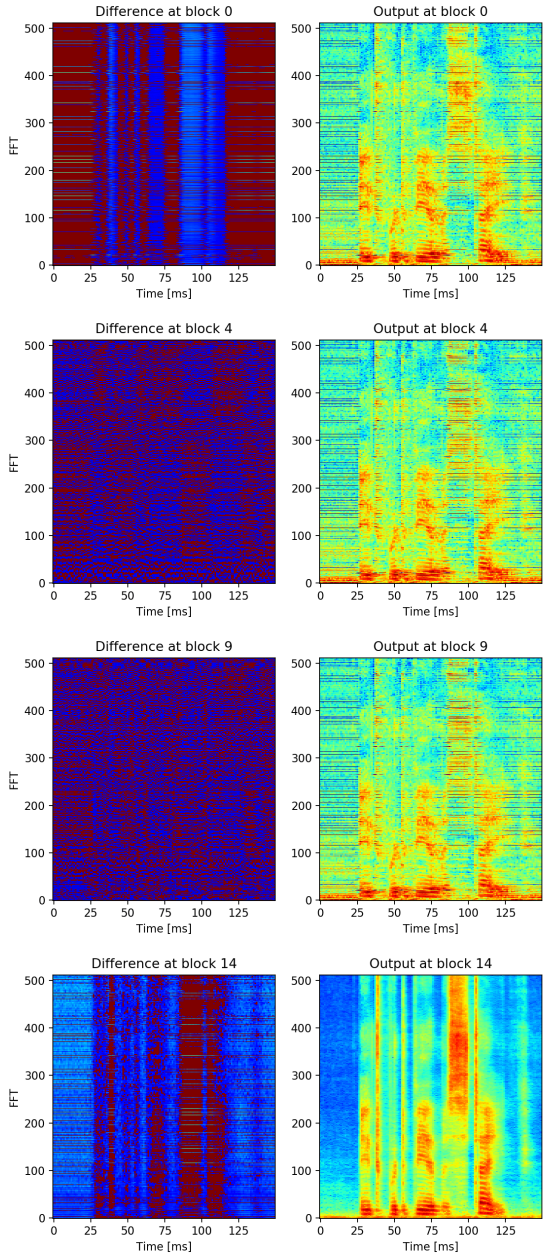


FIGURE 5.6: *Speech enhancement reconstructed output of progressive steps with* CCRN-State *in a signal sample 1.*

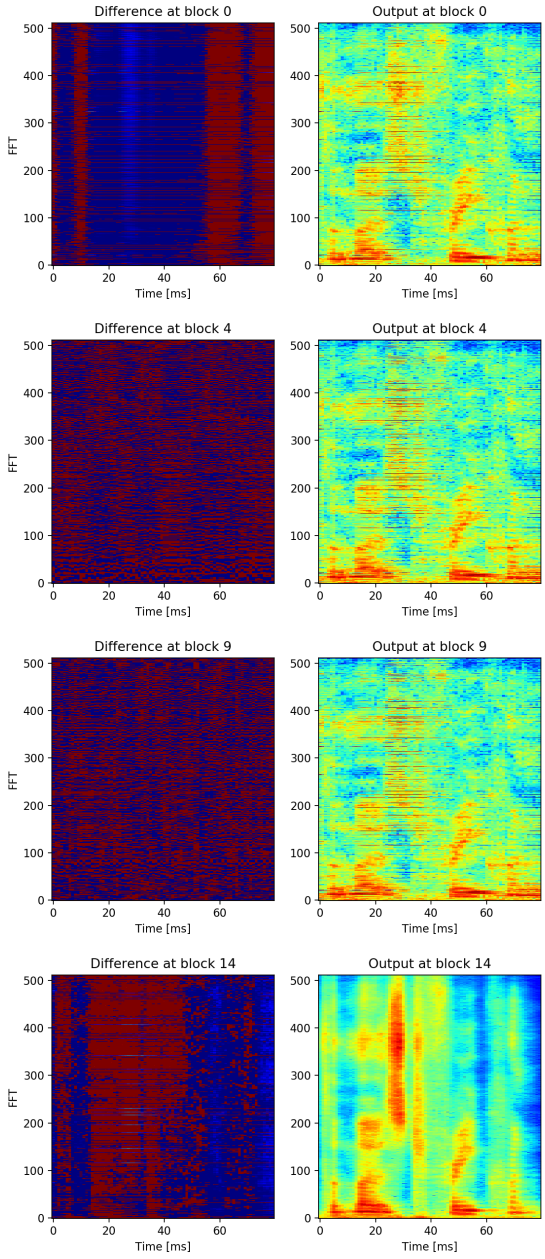


FIGURE 5.7: *Speech enhancement reconstructed output of progressive steps with* CCRN-State *in a signal sample 2.*

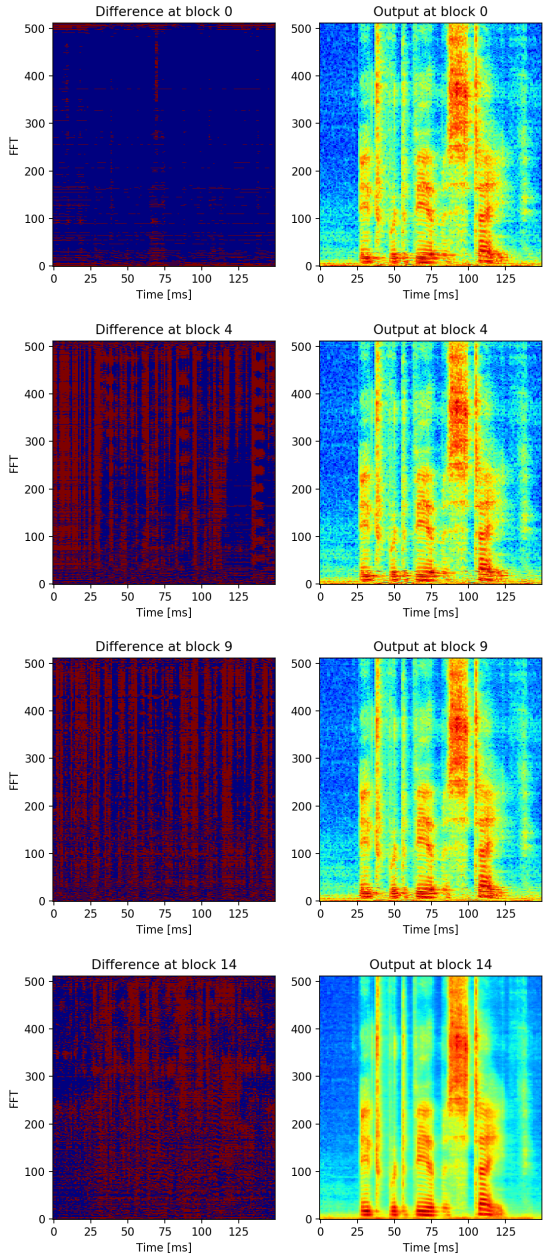


FIGURE 5.8: *Speech enhancement reconstructed output of progressive steps with* CCRN + Progressive Supervision *in a signal sample 1.*

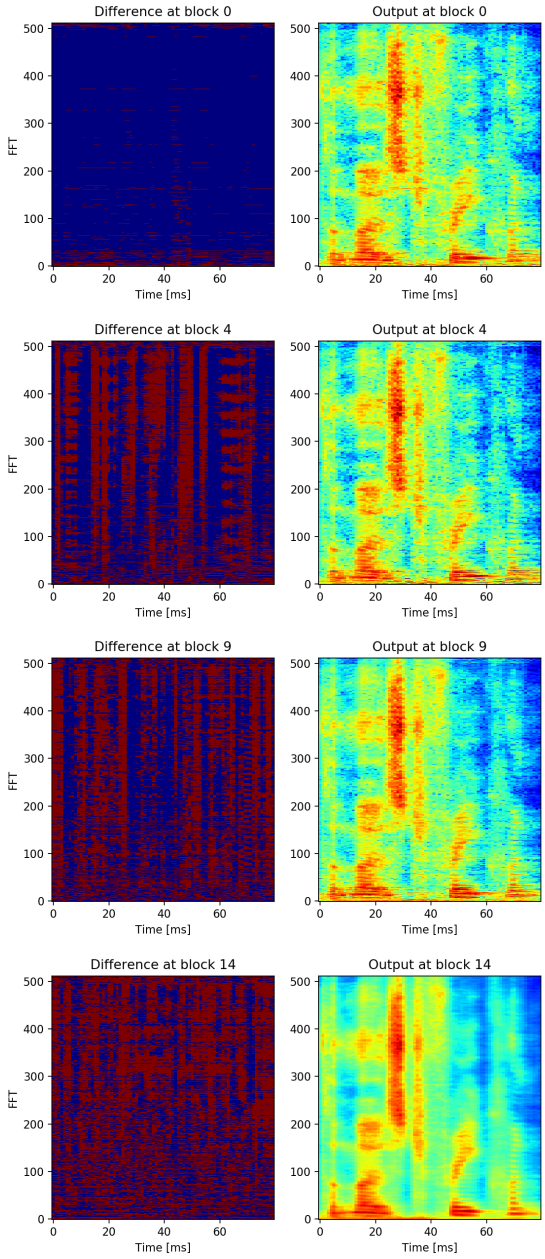


FIGURE 5.9: *Speech enhancement reconstructed output of progressive steps with* CCRN + Progressive Supervision *in a signal sample 2.*

5.8 and 5.9, we can see that each block makes specific changes to improve the speech quality progressively. However, too much smoothing can make the sound feel unnatural.

The visual interpretation of the enhancement process provides insight but does not fully confirm the impact on *SE* performance. The subsequent sections will evaluate the model's accuracy objectively.

5.3 Experimental setup

Training input examples were dynamically generated by distorting contiguous random sequences of 200 samples from the Timit (Garofolo et al., 1993), Librispeech (Panayotov et al., 2015), and Tedlium (Rousseau et al., 2014) databases. The network was trained using the AdamW algorithm (Kingma & Ba, 2015; Loshchilov & Hutter, 2017).

Approaches were evaluated using the official Development and Evaluation sets of the REVERB Challenge (Kinoshita et al., 2013). This dataset includes simulated speech created by convolving the WSJCAM0 Corpus (Robinson et al., 1995) with three different measured Room Impulse Responses (*RIR*) ($RT_{60} = 0.25, 0.5, 0.7s$) at two speaker-microphone distances: near ($0.5m$) and far ($2m$). Stationary noise recordings from the same rooms were added to achieve a Signal-to-Noise Ratio (SNR) of 20 dB. Additionally, the dataset contains real recordings from a reverberant meeting room ($RT_{60} = 0.7s$) at two speaker-microphone distances: near ($1m$) and far ($2.5m$), taken from the MC-WSJ-AV corpus (Lincoln et al., 2005). Real speech samples from VoiceHome v0.2 (Bertin et al., 2016) and v1.0 (Bertin et al., 2019) were also utilized. VoiceHome samples were recorded in a domestic environment from three real homes, incorporating typical household background noises such as dishwashers, vacuum cleaners, and televisions.

Performance was compared with the state-of-the-art dereverberation method known as Weighted Prediction Error (*WPE*) (Nakatani et al., 2010), which is effective in reducing reverberation and enhancing speech quality. The version of *WPE* employed is the more recent DNN-based (Drude et al., 2018) which utilizes an *LSTM* architecture (Kinoshita et al., 2017).

To assess the enhancement quality, distortion reduction was measured using the Log-likelihood ratio (*LLR*) (Loizou, 2011), calculated over the active speech segments. Lower *LLR* values indicate reduced spectral distortion and thus better speech quality. Conversely, the reverberation level of the signal was evaluated using the Speech-to-Reverberation Modulation Energy Ratio (*SRMR*) (Falk et al., 2010a), where higher values suggest improved speech clarity.

5.4 Results and Discussion

5.4.1 Comparative Analysis of Speech Quality Enhancement Methods

Table 5.1 shows how well different methods improved speech quality by reducing distortion. The first row compares the original reverberant speech with the outcomes after applying *WPE* or our *CCRN*-based techniques.

Although all the methods discussed in this chapter improved the clarity of distorted speech, our *CCRN*-based architectures performed better than *WPE* in reducing distortion. Notably, the *CCRN + Progressive Supervision* approach achieved the best results, even though the *CCRN-State* architectures had more flexibility.

TABLE 5.1: *LLR distance in simulated reverberated speech samples from REVERB Dev & Eval datasets.*

Methods	REV-Dev	REV-Eval
Unprocessed	0.63	0.64
WPE (Drude et al., 2018)	0.60	0.60
CCRN	0.52	0.53
+Prog Sup	0.49	0.49
CCRN-State	0.51	0.53
+Prog Sup	0.53	0.54

5.4.2 Speech Dereverberation Across Real and Simulated Environments

Table 5.2 presents the average Speech-to-Reverberation Modulation Energy Ratio (SRMR) scores for both simulated and real speech samples. The first row shows the scores for speech that has not been processed.

The best results across all tested datasets came from using *CCRN + Progressive Supervision*. This top performance is consistent with earlier findings using the LLR metric. Such consistency across different datasets demonstrates the robustness of this method—it works well across various types of speech and noise, not just specific ones.

This method also shows promising results in real environments, not just with simulated data. This suggests it could be effective in practical, everyday situations.

Despite being trained only with artificial reverberation, all *CCRN* models performed well with real-world reverberated speech, indicating they can handle actual reverberation effectively.

TABLE 5.2: *Speech quality through* SRMR *results for real reverberated speech samples.*

Methods	REV-Dev	REV-Eval	VH-v0.2	VH-v1.0
Unprocessed	3.79	3.18	3.19	4.51
WPE (Drude et al., 2018)	4.17	3.48	3.28	4.96
CCRN	4.70	4.11	5.14	5.98
+Prog Sup	5.01	4.44	6.13	7.01
CCRN-State	4.65	3.87	5.09	6.43
+Prog Sup	4.88	4.20	5.35	0.62

Reverberation level, Room sizes, and Near & Far field

Figure 5.10a displays how the Speech-to-Reverberation Modulation Energy Ratio (*SRMR*) scores change as the reverberation level increases in different room sizes, each with specific reverberation times ($RT_{60}(s)$).

All methods based on *CCRN* perform better than the *WPE* baseline. Particularly, the *CCRN + Progressive Supervision* method delivers the best speech quality under all tested conditions.

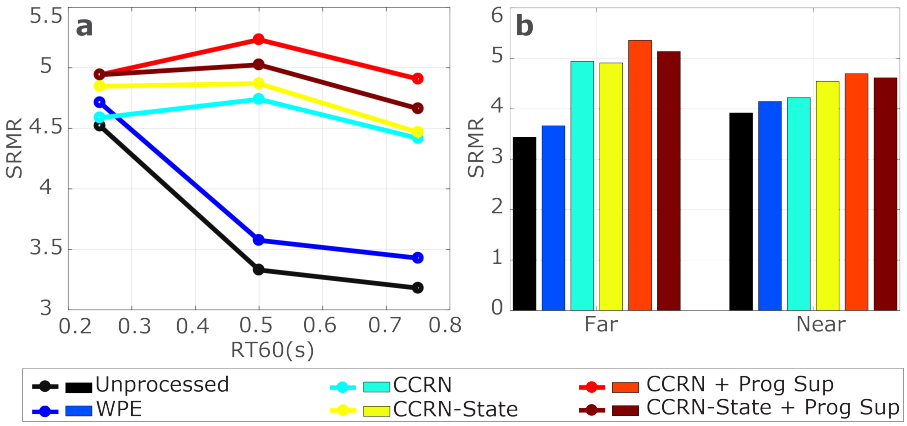


FIGURE 5.10: *Speech quality through SRMR measure in simulated reverberated speech samples from REVERB Dev & Eval datasets.*

This trend is consistent in both "Far" (250 meters) and "Near" (50 meters) conditions, as shown in Figure 5.10b. Notably, *CCRN + Progressive Supervision* improves the results, especially in far-field conditions.

The *Progressive Supervision* technique significantly enhances the network's performance, moving beyond mere frequency adjustments. It helps to fine-tune the network's parameters and progressively shapes the spectrum towards an improved final speech output.

5.4.3 Progressive Supervision in Speech Enhancement Architectures

The enhancement processes in both the *CCRN* and *CCRN-State* architectures produce complex spectral outputs that do not clearly reveal the internal opperation of the networks. It is possible that these representations are simply encodings of the input.

With the introduction of *Progressive Supervision*, the networks begin to demonstrate how the signal improves step-by-step throughout the processing blocks. This cost function also helps regulate the network by discouraging the network from storing auxiliary information different from the spectrum in the residual path, leaving the residual blocks

the task of reconstructing the objective spectrum from the signal at the current stage.

The enhancing patterns at each stage can be viewed as subtracting unwanted components from the spectral power of the signal. Each Residual Block (*RB*) calculates a portion of this power to be subtracted or added, and through the residual connections, these corrections are incrementally applied to the distorted signal being the process fully traceable.

In this framework, the *CCRN + Progressive Supervision* approach shares similarities with the *WPE* method, where both aim to subtract estimated spectral power. However, *CCRN + Progressive Supervision* differs as it performs multiple subtractions, corresponding to the number of blocks in the architecture, using convolutional layers in each *RB*, whereas *WPE* utilizes an *LSTM* for its calculations.

An added benefit of *CCRN + Progressive Supervision* is the interpretability it provides during training and inference operations. We can monitor the reconstruction error at each block, which allows for the efficient training of larger networks. By understanding which blocks contribute to significant improvements, we can optimize the network to operate just with the essential *RBs*, enhancing efficiency especially when processing a clean signal since we could derive from this architecture networks were the number of blocks to process was variable depending the quality of the signal.

5.5 Conclusions

This chapter introduced a deep learning approach for enhancing speech, utilizing residual networks. Through a detailed examination of how the network modifies the log magnitude spectrum step by step, we were able to design an improved architecture aimed specifically at enhancing speech clarity.

The use of "Progressive Supervision" effectively guided the network towards more accurate enhancements by focusing on incremental changes that can be made at each stage to decrease the loss function, allowing the interpretability of the intermediate results in the process. Our

approach has advanced beyond existing methods, finding a good balance between reducing reverberation and minimizing changes to the sound's natural qualities as shown in the objective metrics evaluated.

The insights gained from a detailed analysis of the network's internal processes proved invaluable. They helped us develop more effective methods for improving speech quality.

The next chapter, titled "Progressive Loss for Dereverberation and Noise Reduction," will expand on the concept of progressive loss. It will explore how to adapt this innovative supervision technique also to reduce noise, aiming to refine our models for handling more complex sound environments. This next step will build directly on the foundations laid in this chapter, using the robust framework of residual networks and the benefit of traceability and interpretability of the model.

Chapter 6

Progressive Loss Strategies for Enhanced Speech

6.1 Introduction

Most Deep Neural Network Speech Enhancement (DNN-SE) methods typically operate as a black-box taking a noisy signal and producing an enhanced one. This process is dificult to interprete and organize as classical algorithmic steps. However, speech enhancement can be thought of as a gradual process. Initially, the system performs raw cleaning, then focuses on finer details. This raises a question: Is gradual cleaning useful for better enhancement?

Recently, researchers explored this gradual approach through Progressive Speech Enhancement (PSE) (Gao et al., 2018, 2016; Llombart et al., 2019a). PSE breaks down the learning process into multiple stages, optimizing the target progressively. Each stage's subproblem helps improve the next stage's learning. Previous work shows that PSE often yields better results compared to traditional DNN-SE methods.

In (Gao et al., 2018, 2016), the focus was on improving the Signal-to-Noise Ratio (SNR) in steps. First, a Feed-Forward Deep Neural Network (FF-DNN) used a regression scheme to learn an Ideal Binary Mask, enhancing SNR in 10dB increments (Gao et al., 2016). Later, this approach was extended using LSTM architectures (Gao et al., 2018),

which initially showed performance degradation but eventually improved by incorporating knowledge from previous steps. In (Llombart et al., 2019a), a Wide Residual Network (WRN) was used for step-by-step enhancement, providing insights that helped modify the network architecture for better results. This design calculates the Mean Square Error (MSE) of the Log-Spectral Amplitude (LSA) between the enhanced signal and the reference at each stage, helping to avoid vanishing gradients and enabling interpretability of the process.

Previous results indicate that PSE can improve enhancement performance. This study extends PSE by focusing on its regularization effect when training DNN-SE models. We compare PSE applied to two architectures: Convolutional Neural Network (CNN) and Residual Neural Network (ResNet). CNNs are common in speech technologies but suffer from vanishing gradients in deeper structures. ResNets, an evolution of CNNs, use residual connections to mitigate this problem. We also compare two criteria for progressive loss function optimization: Weighted Progressive (WP) (Llombart et al., 2019a) and a new Uniform Progressive (UP) criterion. The UP criterion treats all blocks' reconstruction errors equally in the final optimization. We evaluate these conditions with simulated and real samples for dereverberation and denoising using the REVERB and VoiceHome corpora.

6.2 Foundational Concepts and Evolution

This work builds on the use of CNN architectures and their evolution into ResNet, as presented in Chapter 5. To adapt these architectures to the progressive paradigm, it is necessary to add additional constraints and modify the loss function. In Chapter 5, we introduced a progressive architecture based on ResNet to understand the enhancement process step by step, employing a visualization probe at each network block to visualize the enhanced signal reconstruction at each stage. The following subsections provide an overview of the architecture design and the modified loss function, which form the foundation of this work.

6.2.1 Architecture

CNN architectures can exploit local patterns in the spectrum from both frequency and temporal domains (Fu et al., 2016; Park & Lee, 2017). Noise and reverberation affect the signal's spectral shape over specific time-frequency areas. The natural structure of speech or distortion patterns can show correlation in consecutive time-frequency bins. CNN-based architectures handle this well, making them suitable for speech enhancement. CNNs can also combine with recurrent blocks to model dynamic correlations among frames (Zhao et al., 2018). Figure 6.1a shows a typical CNN structure with different configurations for convolutional layers, batch normalization, and non-linearities.

Adding residual connections enhances the regularization potential of CNNs (Chen et al., 2017). This architecture, known as Residual Neural Network (ResNet), uses shortcut connections between layers. These connections enable deeper and more complex networks, leading to fast convergence and minimal gradient vanishing. Deeper networks offer more detailed representations of the signal's structure, resulting in more accurate speech enhancement being more expressive. Figure 6.1b illustrates the connection among convolutional blocks in a residual approach.

In (Llombart et al., 2019a), we added a constraint to ResNet, maintaining a constant number of channels in all successive blocks. This allowed output reconstruction and visualization at any internal block improving model robustness. Additionally, this architecture uses a weighted composition of reconstruction errors by block for loss function optimization. Each block performs partial reconstruction, with the next block using a previously enhanced signal representation as input.

6.2.2 Optimization Criteria

In (Llombart et al., 2019a), we proposed a Speech Enhancement (SE) system that reconstructs the Log-Spectral Amplitude (LSA) of a noisy signal. The overlap-add mechanism uses the enhanced logarithmic output spectrum and the phase of the original noisy speech to reconstruct the audio signal. The loss function is the classical Mean Square

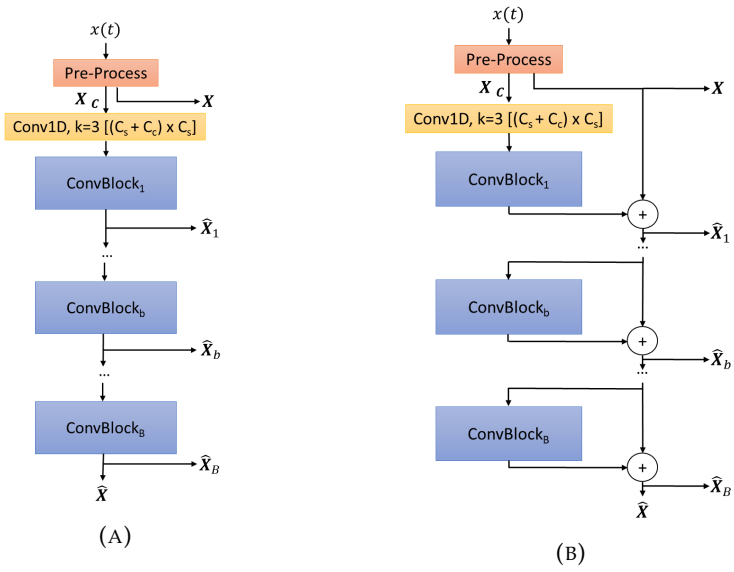


FIGURE 6.1: Architectures presented: (a) Convolutional Neural Network (CNN), (b) Residual Neural Network (ResNet). The convolutional block can have various configurations of convolutional layers and auxiliary layers such as Batch Normalization and non-linearities. The main difference between CNN and ResNet is the residual path in ResNet.

Error (MSE) between the LSA of the reference and the LSA of the enhanced signal, as expressed in Chapter 3 with the equation 3.4. Here, we express it in terms specific to this chapter:

$$MSE(\boldsymbol{y}_{n,\tau}, \hat{\boldsymbol{x}}_{n,\tau}) = \frac{1}{D} \sum_{d=0}^{D-1} (y_{d,n,\tau} - \hat{x}_{d,n,\tau})^2 \tag{6.1}$$

In this equation, D is the signal input dimension. $\boldsymbol{y}_{d,n,\tau}$ and $\hat{\boldsymbol{x}}_{d,n,\tau}$ are the frequency bins of the logarithmic spectrum for training example n and frame τ. $\boldsymbol{y}_{n,\tau}$ is the target vector of the clean LSA reference, and $\hat{\boldsymbol{x}}_{n,\tau}$ is the reconstructed vector of the enhanced signal.

From our experience in previous works (Llombart et al., 2018; Llombart et al., 2019a, 2019b), we learned that using a sequence-based loss function instead of a frame-by-frame loss improves performance. The base loss function is the MSE of the LSA over all examples and the sequence length of an update step as used in Chapters 4 and 5 in equations 4.1 and 5.1. To simplify training, all examples have the same number of frames. This is achieved by randomly cropping the input signals, ensuring that any example selected for training is a random segment of the input.

Finally, (Llombart et al., 2019a) implemented the progressive paradigm by modifying the objective loss function. The MSE between the noisy input LSA and the enhanced LSA is calculated at different network levels or blocks. This progressive loss function is a specific case of the proposal in this thesis. A preliminary experimental study has been presented in the previous chapter which will be expanded with more detail in this chapter.

6.3 Progressive Neural Networks

This chapter explores the potential of the Progressive Speech Enhancement (PSE) paradigm. Previous research has shown that progressive architecture designs can improve SE performance. Building on these findings, we hypothesize that the progressive paradigm not only enhances SE performance but also aids in the regularization of neural

network training. In the following sections, we describe the PSE architecture proposed in this thesis, which is based on the work presented in the previous chapter (Llombart et al., 2019a). Additionally, the study presented in this chapter introduces several novel contributions specifically designed to advance this research.

6.3.1 Architecture

This study explores two DNN architectures: Progressive CNN (P-CNN) and Progressive ResNet (P-ResNet). Building on our previous work (Llombart et al., 2019a) using the ResNet topology, we extend the study to include the CNN topology for comparative purposes and to generalize the progressive paradigm to different architectures.

Figure 6.1 shows the structures of the two architectures under study. The P-CNN and P-ResNet architectures utilize the same convolutional block to ensure comparability. This block starts with a batch normalization stage, followed by a Parametric Rectified Linear Unit (PReLU) non-linearity. Next, a 1D-convolutional layer is applied, maintaining the same number of channels in the output as in the input. This structure is repeated to complete the block, ensuring consistency in the number of channels throughout the architecture. This consistency allows for a reconstructed version of the input using the proposed progressive criteria.

Figure 6.2a depicts the front-end of both architectures. The input signal, $x(t)$, is windowed, and the logarithm of the absolute value of the Short-Term Fourier Transform (STFT) of the input is computed, resulting in the LSA, $\boldsymbol{X}$. Additionally, Mel-Scaled Filter-bank and Mel Frequency Cepstral Coefficients (MFCC) with different windowing processes are obtained to provide extra information to the network, $\boldsymbol{X}_C$.

Maintaining the same number of channels throughout all convolutional blocks ensures that both architectures can produce a consistent enhancement output. This design provides a basis for comparing P-CNN and P-ResNet effectively, ensuring that any differences in performance can be attributed to the architectural variations rather than inconsistencies in the convolutional blocks.

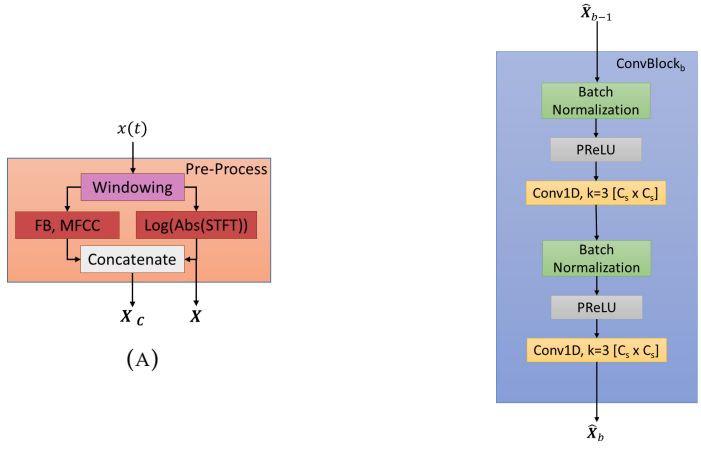


FIGURE 6.2: Structures in P-CNN and P-ResNet: (a) Front-end and (b) Convolutional block.

6.3.2 Progressive Optimization Strategy

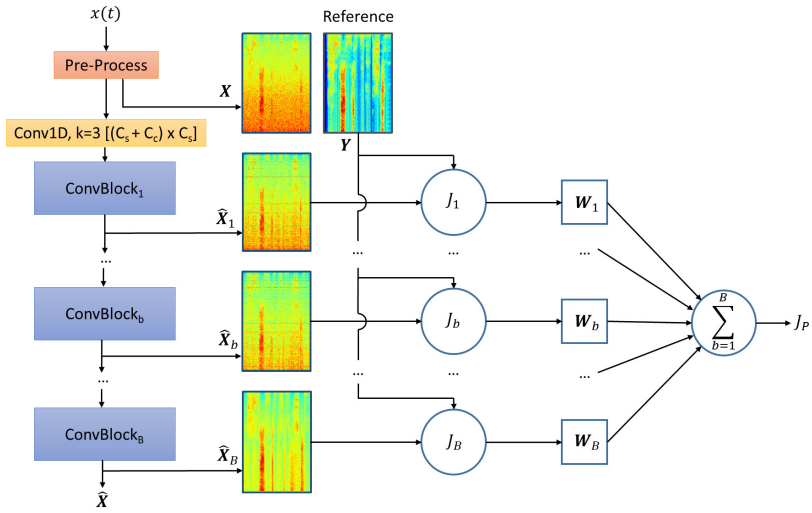


FIGURE 6.3: PSE general architecture for P-CNN and P-ResNet, illustrating the application of the progressive loss that allows direct representation of the output after each block.

In our previous work (Llombart et al., 2019a), we discovered that maintaining the same number of channels as the input signal and achieving full reconstruction through the loss function allows us to track the enhancement progress through the architecture stages during training. After each convolutional block, we minimize the MSE between the clean reference $\boldsymbol{Y}$ and the block output $\hat{\boldsymbol{X}}_b$ (Fig. 6.3). The general definition of the progressive loss function as a weighted sum over the reconstruction loss of each convolutional block is given by:

$$J_P(\boldsymbol{Y}, \hat{\boldsymbol{X}}) = \sum_{b=1}^{B} W_b \cdot J(\boldsymbol{Y}, \hat{\boldsymbol{X}}_b). \tag{6.2}$$

By adjusting the weights in Equation (6.2), we can define different progressive loss function criteria. In (Llombart et al., 2019a), we proposed the Weighted Progressive (WP) loss function. Based on the general definition in Equation (6.2), we describe the WP criterion and introduce the Uniform Progressive (UP) criterion. In the following sections, both criteria are experimentally evaluated with P-CNN and P-ResNet architectures.

6.3.3 Weighted Progressive (WP)

The WP loss function primarily focuses on the final cost, typical in approximation tasks. The cost of all the architecture blocks is uniformly distributed and added in a weighted sum:

$$J_{WP}(\boldsymbol{Y}, \hat{\boldsymbol{X}}_B) = J(\boldsymbol{Y}, \hat{\boldsymbol{X}}_B) + \alpha \frac{1}{B} \sum_{b=1}^{B} J(\boldsymbol{Y}, \hat{\boldsymbol{X}}_b) \tag{6.3}$$

Here, B is the number of blocks in the architecture. Equation (6.3) is a specific case of the general progressive loss function in Equation (6.2), where $W_b = \alpha/B$ for $b = 1, ..., B-1$ and $W_B = 1 + \alpha/B$. This loss function implements progressive processing along blocks, where every intermediate block reconstructs the enhanced signal. This design promotes a progressive enhancement process, transforming from initial to detailed cleaning. In addition it has the benefit of helping the gradient

propagation in deeper architectures and complements the traditional gradient back-propagation across the entire architecture.

6.3.4 Uniform Progressive (UP)

The UP loss function proposes a uniform distribution of the block losses across the architecture:

$$J_{UP}(\boldsymbol{Y}, \hat{\boldsymbol{X}}_B) = \frac{1}{B} \sum_{b=1}^{B} J(\boldsymbol{Y}, \hat{\boldsymbol{X}}_b), \tag{6.4}$$

This is a special case of Equation (6.2) where $W_b = \alpha / B$ for $b = 1, ..., B$. With this strategy, all outputs have the same impact on the reconstruction. Each block equally contributes to the final loss, ensuring that the entire architecture makes a uniform effort in signal reconstruction.

6.4 Experimental setup

This section presents the experimental procedure and data used in this study. We employ the REVERB and VoiceHome datasets for testing. Our evaluation focuses on dereverberation and noise reduction in both simulated and real samples through various experiments.

6.4.1 Training Data

For DNN training, we have used three different public datasets: Tedlium (Rousseau et al., 2014) from Ted talks; Librispeech (Panayotov et al., 2015), audio-books; and Timit (Garofolo et al., 1993), a phonetically ballanced distributed read speech. These datasets are fully employed, without any partition. See Table 6.1 for the characteristics of the datasets.

TABLE 6.1: Training Datasets Description

Dataset	Timit	Librispeech	TedLium
Files	6299	292329	56704
Speakers	630	2484	698
Speech type	Read speech		Conference
Interface	Close Microphone		Auditorium microphone

6.4.2 Data augmentation: Reverberated and Noisy training data

Data augmentation using reverberation and additive noise was applied to the training set. For each random training example, we performed three transformations:

1. **Impulse Responses**: We simulated random rooms and source-receiver distances using Room Impulse Responses (RIR) generated with the Python package rir-generator[1] (Allen & Berkley, 1979b). Table 6.2 details the characteristics of the RIRs used. During the data augmentation loop, we simulated three types of rooms: small, medium, and large, selected with probabilities of 0.5, 0.3, and 0.2, respectively with random sizes and source and microphone position according to the table ranges.

2. **Additive Noise**: Noise was added with a Signal-to-Noise Ratio (SNR) uniformly sampled from $SNR \sim U(5, 25)$ dB, using the Musan dataset (Snyder et al., 2015). We included music and noise files but excluded speech files. Table 6.3 describes the noise characteristics. Among the noise files, crowd noise is included, but no intelligible speech is present.

3. **Time Scaling**: We randomly selected a scale between 0.8 and 1.2. Some signals were unscaled (original speed), while others were either slowed down or sped up.

TABLE 6.2: RIR for training data augmentation.

	Room Impulse Responses		
	Small	Medium	Large
Probability	0.5	0.3	0.2
Size (x,y,z)[m]	$x \sim U(1,6)$ $y \sim U(1,6)$ $z \sim U(2,3.5)$	$x \sim U(6,10)$ $y \sim U(6,10)$ $z \sim U(3,5)$	$x \sim U(10,20)$ $y \sim U(10,20)$ $z \sim U(4,6)$
RT_{60}[s]	$RT_{60} \sim U(0.1, 0.25)$		
Distance[m]	0.5, 1.0, 1.5, 2.0, 2.5		
Microphone type	bidirectional, hypercardiodid, cardioid subcardoid, omnidirectional		

[1] https://github.com/Marvin182/rir-generator

TABLE 6.3: Noise for training data augmentation.

	Noise
Music	659 files
Noise	929 files
SNR [dB]	$SNR \sim U(5, 25)$

6.4.3 Evaluation Data

For evaluation purposes, we use two databases: REVERB (Kinoshita et al., 2013) and VoiceHome (v0.2 (Bertin et al., 2016) and v1.0 (Bertin et al., 2019)). REVERB consists of a development set (REVERB-Dev) for intermediate evaluations and an evaluation set (REVERB-Eval) for confirming results and evaluating the system. VoiceHome evaluates the system in a realistic domestic environment with noise and reverberation. These databases allow us to assess two conditions:

Simulated Data The REVERB dataset provides simulated conditions with speech samples from the WSJCAM0 corpus (Robinson et al., 1995). It includes three types of Room Impulse Responses (RIRs): small, medium, and large rooms with reverberation times of $RT_{60} = 0.25, 0.5$, and $0.7s$, respectively. Each room has two source-microphone distances: near $(0.5m)$ and far $(2m)$. Additionally, stationary noise at $SNR = 20dB$ was added from the same rooms. While the dataset provides eight channels, we only use the first channel for this study. We further augmented the signals with five types of noise at varying SNR levels $(0, 5, 10, 15, 20$, and $25dB)$, including babble noise, café environment noise, music, street traffic noise, and noise from inside a moving tram.

Real Data We use two evaluation sets with real conditions: the real part of REVERB and the VoiceHome datasets (v0.2 and v1.0). The real part of REVERB was recorded in a meeting room with $RT_{60} = 0.7s$ at two distances: near $(1m)$ and far $(2.5m)$, sourced from the MC-WSJ-AV corpus (Lincoln et al., 2005). The VoiceHome dataset reflects a realistic domestic environment with everyday noises such as a vacuum cleaner, dishwashing, and TV sounds.

6.4.4 Speech Quality Measures

To evaluate the denoising and dereverberation performance of the Progressive Speech Enhancement (PSE) method, we use several key metrics. As discussed in Chapter 3, we measure the segmental Signal-to-Noise Ratio (SNR) (Kim & Stern, 2008) and the Speech-to-Reverberation Modulation Energy Ratio (SRMR) (Falk et al., 2010b; Santos et al., 2014). Higher values in these metrics indicate better speech quality, as they reflect reduced noise and reverberation.

However, speech enhancement can introduce distortion to the output speech. To assess this, we measure the distortion between the clean reference and the enhanced speech using the Log-Likelihood Ratio (LLR) (Loizou, 2011). Lower LLR values indicate less distortion and thus better speech quality. Balancing noise/reverberation reduction and minimizing distortion is crucial for a comprehensive assessment of the SE method's performance. The optimal enhancement system improves SNR or SRMR while keeping LLR as low as possible.

Other commonly used measures in speech enhancement include the Perceptual Evaluation of Speech Quality (PESQ), Perceptual Objective Listening Quality Analysis (POLQA), and Short Term Objective Intelligibility (STOI). These metrics aim to estimate speech quality perceptually. However, they can be complex, difficult to use, and sometimes not publicly available.

Recent studies suggest a high correlation between perceptual speech quality and commonly used measures such as SNR, SRMR, and LLR (Gelderblom et al., 2018; Santos & Falk, 2019). These simpler measures often correlate well with more complicated metrics like PESQ and STOI. Therefore, in this work, we focus on SRMR, SNR, and LLR to provide clear and reliable results.

6.4.5 Neural Network Configuration

The input for the CNN, ResNet, P-CNN, and P-ResNet architectures is the logarithm of the magnitude of the 512-point Short-Time Fourier

Transform (STFT) of the corrupted signal sampled at 16 kHz, computed every 10 ms using a 25 ms sliding Hamming window. Additionally, we concatenate the Mel-Scaled Filter-bank and the Mel Frequency Cepstral Coefficients (MFCC) as auxiliary inputs, providing different frequency and temporal resolution views with 32, 50, and 100 frequency bins, respectively. These are computed every 10 ms with sliding Hamming windows of 25 ms, 50 ms, and 75 ms.

For all experiments, we use the AdamW optimizer with 0.001 learning rate and $5e^{-5}$. Each layer contains 512 neurons, adhering to the principle of maintaining a consistent number of channels throughout the architecture. The training process runs for 900 epochs, each consisting of $10,000$ input files randomly selected from the training set, ensuring each file is used once before repeating.

For the J_{WP} loss function, we set $\alpha = 0.1$, as in (Llombart et al., 2019a), which provided the best SRMR value on the REVERB-Dev dataset.

6.5 Preliminary gradient study

This section presents a preliminary study on gradient behavior to explore the hypothesis that the progressive paradigm aids in training regularization by addressing vanishing gradient problems. When gradients back-propagate through many layers, they often lose energy, reducing their ability to adjust the weights of layers close to the input. The proposed PSE method injects a fresh and stronger gradient after each block, helping to move the weights of each layer effectively.

We designed an experiment to observe the gradient energy that modifies the weights of the first convolutional block during the first 100 optimization updates. This procedure was repeated 100 times with different weight initializations to observe the variance among different starts and the variation of gradient energy during optimization.

Figure 6.4 shows the results for P-CNN and P-ResNet architectures, comparing non-progressive baselines with each proposed progressive procedure. There is a noticeable difference in gradient behavior between the two structures. In P-CNN, there is a significant difference in gradient energy among the compared systems. The lowest energy

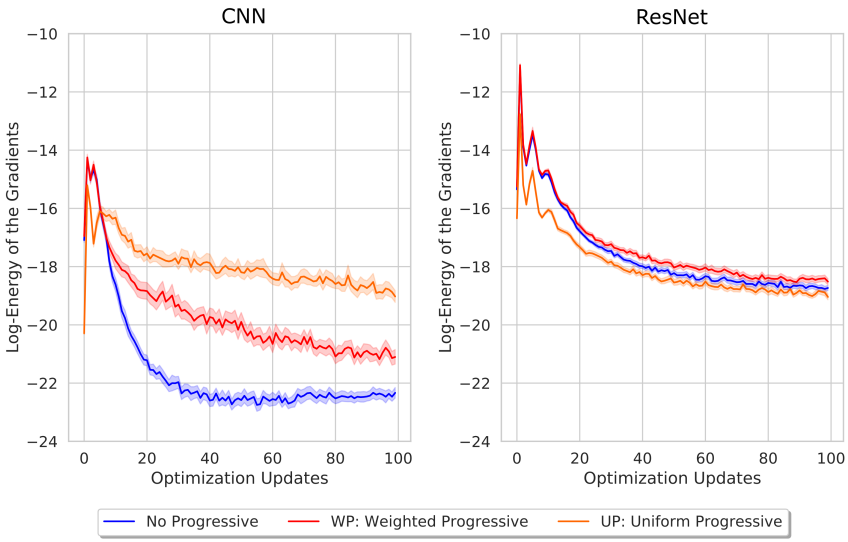


FIGURE 6.4: Mean and variance (shaded area) of the log-energy of the gradients from 100 random network initializations during the first 100 mini-batches of training. The log-energy is measured on the nearest block to the input.

corresponds to the baseline architecture without any progressive assumption. In contrast, the progressive mechanisms show a significant increase in gradient energy. These boosted gradients are more effective adjusting the weights, allowing better optimization of the entire architecture.

In contrast, P-ResNet shows no relevant difference between the gradient energy of the progressive techniques and the non-progressive baseline at the first convolutional block. P-ResNet is designed to handle the vanishing gradient problem, and thanks to residual connections, the gradients can propagate to the first layers without vanishing. In this case, injecting new gradients does not significantly enhance the existing gradients. However, the new gradients are more accurate because they come directly from the target evaluation at the output of each architecture block.

6.6 Results and discussion

6.6.1 Architecture depth analysis

TABLE 6.4: Speech quality in terms of SRMR for simulated and real reverberated speech samples through architecture depth for REVERB-Dev dataset. The last rows represent the mean and standard deviation along the experiments presented in each column.

Archictecture	Blocks Depht	Simulated	Real	AVG±STD
CNN	8	7.33	6.05	6.69±0.64
	16	7.60	5.98	6.79±0.81
	24	8.87	4.76	6.81±2.05
	32	7.01	3.35	5.18±1.83
ResNet	8	8.23	6.82	7.52±0.70
	16	8.27	5.81	7.04±1.23
	24	8.14	5.77	6.97±1.16
	32	8.56	6.33	7.44±1.11
P-CNN with WP	8	6.49	4.90	5.69±0.79
	16	8.96	3.74	6.35±2.61
	24	6.18	2.07	4.12±2.05
	32	7.65	2.33	4.99±2.66
P-CNN with UP	8	7.53	6.32	6.92±0.60
	16	7.70	7.26	7.48±0.22
	24	8.09	6.90	7.49±0.59
	32	7.41	6.34	6.87±0.53
P-ResNet with WP	8	8.31	7.06	7.68±0.62
	16	8.41	7.14	7.77±0.63
	24	8.03	6.53	7.28±0.75
	32	7.98	5.97	6.97±1.00
P-ResNet with UP	8	7.91	6.91	7.41±0.50
	16	8.05	6.85	7.45±0.60
	24	8.02	6.91	7.46±0.55
	32	7.78	6.62	7.20±0.58

Progressive SE methods use multiple steps to enhance the signal, requiring us to determine the optimal number of blocks for the architecture. Table 6.4 shows the architecture depth study in terms of SRMR over the REVERB-Dev dataset, presenting results for both simulated and real conditions, as well as their average.

Results indicate that the configuration with 16 blocks achieves the best performance across all evaluated conditions. Progressive systems demonstrate high SRMR for both simulated and real conditions, showcasing the consistency and better generalization capability of the progressive strategy in DNN training.

For the CNN topology, the reference system's performance in real conditions degrades quickly with increased architecture depth. However, the P-CNN with the Uniform Progressive (UP) criterion performs better than the CNN reference system, indicating that P-CNN with UP does not degrade as rapidly as the reference system as depth increases.

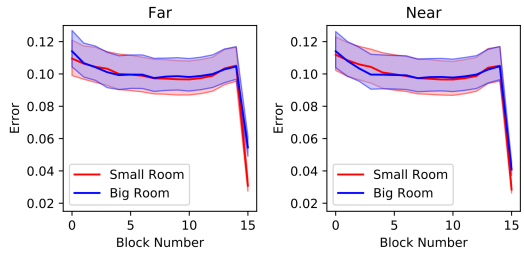
For the ResNet topology, residual connections effectively support larger block configurations. For instance, the ResNet reference system achieves the best performance in simulated conditions with a deeper architecture (32 blocks). However, in real conditions, the ResNet reference system performs best with 8 blocks. Notably, the P-ResNet with the Weighted Progressive (WP) criterion surpasses the reference system's best result in real conditions with 16 blocks, which is also the optimal configuration for P-ResNet in simulated conditions.

6.6.2 Progressive enhancement along architecture blocks

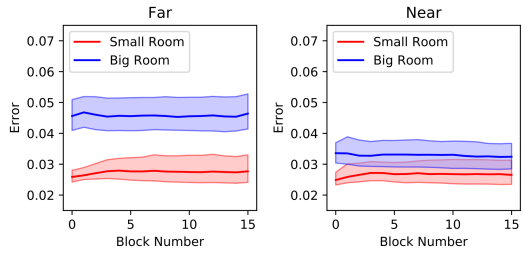
In this section, we analyze the behavior of the Progressive Speech Enhancement (PSE) method on speech data affected by different reverberation levels. We use signals from large and small rooms in the simulated condition of REVERB-Dev, which provides samples with varying room sizes and source-microphone distances.

Figure 6.5 shows the evolution of the Mean Square Error (MSE) between the clean reference and the reconstruction at each block output for P-CNN and P-ResNet with Weighted Progressive (WP) and Uniform Progressive (UP) supervisions.

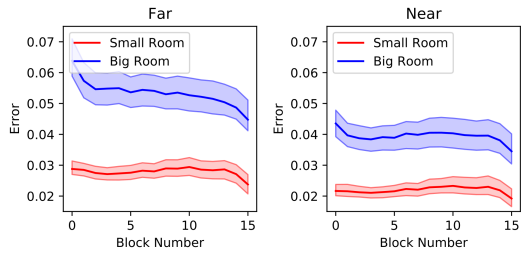
Firstly, we observe that the reconstruction error decreases with the proximity between the source and the microphone, resulting in less error for near samples compared to far samples. In near conditions, the direct path speech energy is higher than the reverberant path energy, reducing the impact of reverberation and thereby lowering the error during evaluation.



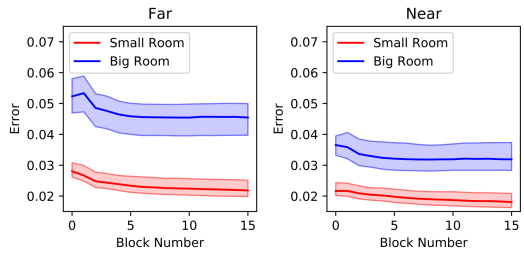
(A) P-CNN with WP



(B) P-CNN with UP



(C) P-ResNet with WP



(D) P-ResNet with UP

FIGURE 6.5: MSE between clean reference and reconstruction output at each block of the different architectures and progressive methods on the REVERB-Dev set. The dark line shows the mean, and the shaded area between Q1 and Q3 shows the variability of the MSE of the examples on the set.

Regarding room size, the far distance in a large room yields the highest errors for all evaluated cases, which is expected due to the higher reverberation level in this condition. However, in small rooms, there is no significant difference between far and near conditions because the reverberation level is generally lower.

For progressive supervision, P-CNN with WP shows a noticeable drop in error at the last block. In P-ResNet with WP, there is also a drop at the last block, but the error reduction is more evenly distributed across all blocks. WP focuses significant reconstruction effort on the final block. Conversely, UP distributes the reconstruction effort more gradually through the blocks. In P-CNN with UP, the error remains relatively stable across all blocks. In the small room condition, the error initially increases at the first block before stabilizing, suggesting that improving SE performance might benefit from early layer reconstruction. In the large room condition, the error decreases consistently across the blocks. P-ResNet with UP shows a constant decrease in error through the blocks as expected.

Results indicate that progressive supervision benefits the SE system, although the effectiveness of a particular strategy may vary with the architecture. Overall, we conclude that PSE contributes positively to neural network regularization.

6.6.3 Dereverberation

To assess the impact of the PSE proposal in dereverberation tasks, we use the SRMR quality measure and LLR to evaluate the distortion introduced by the method (the latter only for simulated conditions). Experiments are conducted on REVERB-Eval and VoiceHome v0.2 and v1.0 datasets, which also include some noisy conditions. For comparison, we use a DNN variation of the state-of-the-art dereverberation method, Weighted Prediction Error (WPE) (Nakatani et al., 2010), enhanced with Long Short-Term Memory (LSTM) cells (Drude et al., 2018).

Table **??** presents the SRMR and LLR results for both reference and progressive systems. PSE methods show the best results. In simulated conditions, P-CNN with WP achieves the highest SRMR but also

introduces more distortion. Conversely, P-ResNet with WP achieves slightly lower SRMR but with significantly less distortion, offering a better trade-off for speech quality. We conclude that while PSE introduces some additional distortion, it is not significant compared to the improvement in SRMR.

TABLE 6.5: Speech quality in terms of SRMR and LLR for simulated and real reverberated speech. The last row represents the mean and standard deviation along the experiments presented in each column. Dark gray corresponds with the best dataset value, light gray shows the second best value.

		REVERB Eval Simulated	REVERB Eval Real	Voice Home V0.2	Voice Home V1.0	AVG SRMR
Unproc.	SRMR	6.34	3.44	3.23	4.04	4.26±1.24
	LLR	-	-	-	-	-
WPE	SRMR	6.64	3.74	3.38	4.47	4.56±1.26
	LLR	0.57	-	-	-	-
CNN	SRMR	7.37	5.86	5.80	5.89	6.23±0.66
	LLR	0.49	-	-	-	-
ResNet	SRMR	7.90	5.79	5.69	6.13	6.38±0.89
	LLR	0.47	-	-	-	-
P-CNN with WP	SRMR	8.16	3.84	2.58	2.78	4.34±2.26
	LLR	0.79	-	-	-	-
P-CNN with UP	SRMR	7.46	7.23	5.49	5.81	6.50±0.86
	LLR	0.53	-	-	-	-
P-ResNet with WP	SRMR	8.08	7.00	7.32	7.31	7.43±0.40
	LLR	0.48	-	-	-	-
P-ResNet with UP	SRMR	7.84	6.83	5.72	6.27	6.66±0.78
	LLR	0.49	-	-	-	-

In real conditions, the best result for the REVERB dataset is achieved by P-CNN with UP, whereas for the VoiceHome dataset, the best result is obtained by P-ResNet with WP. P-ResNet with WP is the most consistent across both datasets; although it is not the top performer for REVERB, it is a close second. P-CNN with UP shows high variability between simulated and real conditions, possibly due to over-fitting to simulated conditions.

Table **??** also shows the average (AVG) and standard deviation (STD) of the evaluated systems for each architecture. P-ResNet with WP achieves the best result with the least variability across evaluation datasets. This outcome demonstrates that P-ResNet with WP is the most regularizing structure and the most general-purpose architecture for dereverberation tasks.

6.6.4 Noise reduction in reverberant environment

This section discusses the performance of the proposed systems on noise reduction using the noisy simulated data on REVERB (see section 6.4.3). The SNR measures the speech quality performance of SE for denoising level, and LLR, for distortion level.

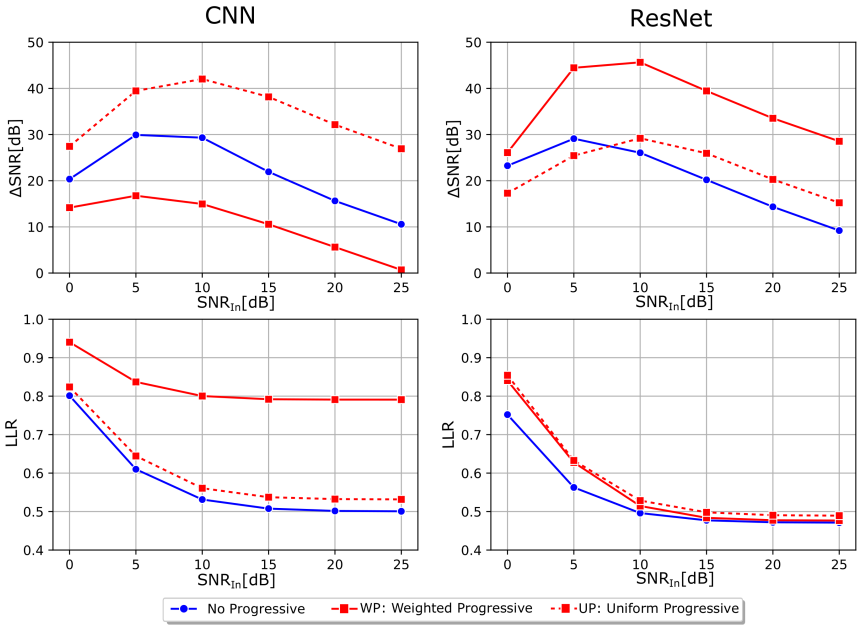


FIGURE 6.6: $\Delta SNR = SNR_{Out} - SNR_{In}$ and LLR after enhancement for both architectures in REVERB-Eval with noise.

Figure 6.6 illustrates the SNR increase (ΔSNR) and Log-Likelihood Ratio (LLR) after speech enhancement (y-axis) versus the initial SNR at the input (x-axis). ΔSNR represents the improvement in the estimated

TABLE 6.6: Summary of speech quality in terms of ΔSNR and LLR for simulated reverberated and noisy speech samples in REVERB-Eval. Mean through all noise types and initial SNR levels conditions evaluated.

	Reference Systems		Progressive Systems			
	CNN	ResNet	CNN with WP	CNN with UP	ResNet with WP	ResNet with UP
ΔSNR $[dB]$	21.29	20.36	10.46	34.36	36.28	22.23
LLR	0.58	0.54	0.83	0.61	0.57	0.58

output SNR, calculated using the Wada method (Kim & Stern, 2008), compared to the input SNR: $\Delta SNR = SNR_{Out} - SNR_{In}$.

The results align with the dereverberation findings discussed in Section 6.6.3. For the CNN topology, P-CNN with UP achieves the best outcome, while for the ResNet topology, P-ResNet with WP delivers the best performance.

During evaluation, we used input signals with an SNR of 0, which were not included in the training process. Despite this, all systems performed well in enhancing these signals. Additionally, as the input SNR improves, the systems' performance increases until the input is so clean that further enhancement is minimal.

In terms of distortion, systems without progressive supervision exhibit lower LLR values, although P-ResNet systems are close to them. In the CNN architecture, UP does not significantly increase distortion compared to its reference system. However, in the ResNet architecture, all systems introduce similar levels of distortion, with the reference system causing the least distortion at low input SNR levels.

Table 6.6 summarizes the noise reduction evaluation results, showing the average ΔSNR and distortion for all noise types and initial SNRs for each evaluated system (detailed results are available in Appendix B). The best denoising system is P-ResNet with WP, followed by P-CNN with UP. These systems significantly outperform their respective reference architectures.

Considering the best trade-off between SNR improvement and distortion, the reference ResNet introduces the least distortion but performs poorly in denoising. The second-best system in terms of distortion is P-ResNet with WP, which also excels in denoising tasks. Therefore, we conclude that the progressive strategy is effective for noise reduction, with P-ResNet with WP offering the best balance between SNR improvement and minimal distortion.

6.7 Conclusions

This study investigated the Progressive Speech Enhancement (PSE) method using both CNN and ResNet architectures. We explored two criteria for progressive loss function optimization: the Weighted Progressive and the Uniform Progressive strategies, with the latter being a novel proposal. The results demonstrated that progressive supervision is valuable for regularization in both CNN and ResNet architectures.

The PSE method effectively achieves regularization in dereverberation and denoising tasks without significantly increasing distortion. Among the architectures studied, P-ResNet with Weighted Progressive showed the most consistent performance across various conditions, providing a positive trade-off in terms of result quality. This architecture remained competitive in all experiments, making it a reliable choice for speech enhancement tasks.

Overall, the architectures evaluated in this study, particularly P-ResNet with Weighted Progressive, achieved excellent results in both dereverberation and denoising, proving their suitability for speech enhancement applications.

Chapter 7

Conclusions

In this thesis, we aimed to improve speech enhancement using deep neural networks, focusing on developing and evaluating new methods that surpass existing techniques. The main contributions have been to introduce the Wide Residual Network (WRN) architecture for speech enhancement, enhance model interpretability, and explore progressive loss strategies.

Our research demonstrated that the WRN architecture significantly outperforms traditional methods in various noisy environments. Additionally, visualization techniques provided valuable insights into neural network behavior, leading to the development of effective progressive loss strategies that improve speech quality.

The findings contribute to the field by offering a robust architecture for speech enhancement that balances performance and computational complexity. Our interpretability enhancements help bridge the gap between 'black box' models and user understanding, providing more trust in neural network applications.

7.1 WRN-based Speech Enhancement

In Chapter 4, WRN-based Speech Enhancement introduced a novel approach to speech enhancement using a Wide Residual Network (WRN) architecture. By leveraging the powerful representations provided by a wide topology of Convolutional Neural Networks (CNNs) with residual connections, this method demonstrated significant improvements

over existing state-of-the-art methods, particularly in handling far-field reverberated speech. The following points summarize the key findings and insights:

- **Superior Performance:** The WRN method significantly outperformed the state-of-the-art RNN-LSTM-based Weighted Prediction Error (WPE) method, particularly in handling far-field reverberated speech across various room sizes. These promising results and methodological improvements provide a strong foundation for further research, ensuring the ongoing relevance and effectiveness of the WRN method in speech enhancement.
- **Benefits of Residual Connections:** The inclusion of residual connections allowed the network to maintain linear signal pathways while enhancing non-linear corrections, improving practical application performance.
- **Visualization for Enhanced Interpretability:** Visualization techniques in subsequent chapters help to understand the WRN model's enhancement process, addressing the "black box" nature of deep learning and identifying key steps for optimization.

In conclusion, the WRN method presented in this chapter offers a robust and effective solution for speech enhancement, outperforming existing methods in challenging conditions. The subsequent chapters provide detailed analyses and methodological improvements that enhance both the performance and interpretability of the model. These advancements ensure that the WRN method remains a solid foundation for further research and development in the field of speech enhancement, culminating in the publication of these findings in:

- **J. Llombart**, D. Ribas, A. Miguel, L. Vicente, A. Ortega, and E. Lleida, **"Speech Enhancement with Wide Residual Networks in Reverberant Environments"** Proc. Interspeech 2019, 2019, pp. 1811–1815.

7.2 Visualization Techniques in Speech Enhancement

In Chapter 5, Visualization is focused on enhancing the interpretability of speech enhancement through deep learning architectures. The exploration of new architectures aimed to make the enhancement process more transparent and understandable. By employing various visualization techniques, the chapter aimed to demystify the "black box" nature of neural networks and improve the overall design and effectiveness of speech enhancement solutions. Outlined below are the primary findings and insights:

- **Proposed Architectures:** The chapter introduced new architectures, including the Constant Channel Residual Network (CCRN) and the Constant Channel Residual Network with State Path (CCRN-State), which improved speech enhancement performance and interpretability. These architectures showed promising results in enhancing speech quality and provided valuable insights into the models' internal workings, suggesting a good trade-off between performance and interpretability.
- **Visualization Techniques:** Visualization probes were employed to monitor and understand the enhancement process at each network block. This step-by-step supervision provided insights into how different network components contribute to the overall enhancement.
- **Progressive Supervision:** The concept of progressive supervision was implemented to track the enhancement process incrementally, which helped to identify critical stages that significantly impact speech quality.

In conclusion, we successfully demonstrated that integrating visualization techniques into deep learning architectures for speech enhancement can significantly enhance the interpretability and performance of the models. The insights gained from this study provide a solid foundation for further research and development in creating more transparent and effective speech enhancement solutions. The findings of this chapter have been published in:

- **J. Llombart**, D. Ribas, A. Miguel, L. Vicente, A. Ortega, and E. Lleida, **"Progressive Speech Enhancement with Residual Connections"** Proc. Interspeech 2019, 2019, pp. 3193–3197.

7.3 Progressive Loss Strategies in Speech Enhancement

In Chapter 6, We explored progressive loss strategies for enhanced speech using neural network architectures. The aim was to improve speech enhancement by refining the training process and optimizing performance through innovative loss strategies. The following key points summarize the conclusions drawn:

- **Progressive Loss Strategies:** The implementation of progressive loss strategies, including Weighted Progressive (WP) and Uniform Progressive (UP) methods, significantly improved the model's ability to enhance speech quality by systematically reducing noise and reverberation throughout the network.
- **Gradient Analysis:** Preliminary gradient studies indicated that progressive loss strategies help in stabilizing the training process of deep architectures, leading to more consistent and robust performance across different noise conditions and environments.
- **Architecture Depth Analysis:** Deeper neural network architectures, when combined with progressive loss strategies, showed enhanced performance in speech quality measures, indicating the importance of depth in achieving superior results.

In conclusion, this highlighted the effectiveness of progressive loss strategies in enhancing speech quality. These strategies provide a robust framework for further advancements in speech enhancement technology. The findings presented underscore the potential for progressive loss strategies to set new benchmarks in the field, contributing valuable insights for future research and practical implementations. The results of this study have been published in:

- **J. Llombart**, D. Ribas, A. Miguel, L. Vicente, A. Ortega, and E. Lleida, **"Progressive loss functions for speech enhancement with**

deep neural networks" EURASIP Journal on Audio, Speech, and Music Processing, vol. 2021, pp. 1–16, 2021.

7.4 Future Lines of Research

Building on the promising findings and insights from this thesis, several future research directions can be pursued to further advance the field of speech enhancement using deep neural networks. The following points outline potential areas for future exploration and development:

- **Optimization of WRN Architectures:** Further refinement and optimization of Wide Residual Network (WRN) architectures could lead to even greater performance improvements. This includes exploring different configurations and hyperparameters to enhance their ability to handle various noisy environments more effectively.
- **Advanced Visualization Techniques:** Expanding on the visualization methods introduced in Chapter 5, future work could develop more sophisticated tools to provide deeper insights into the neural network's enhancement processes. This could involve real-time visualization techniques that allow for dynamic monitoring and adjustment during the training phase, for example learning the optimal number of denoising steps in the progressive framework.
- **Integrating Additional Deep Learning Models:** Incorporating other advanced deep learning models, such as transformers (Vaswani et al., 2017), state-space models (Gu & Dao, 2023), new recurrent models (Beck et al., 2024), self-supervised learning frameworks (Chen et al., 2022) and Diffusion-based generative models (Ho et al., 2020), could provide complementary strengths to WRN architectures. This integration has the potential to further improve speech enhancement, especially in more challenging acoustic environments, but some methods may generate new speech not present in the source audio, so preventing this is one of the main lines of investigation.

- **Exploration of Progressive Loss Strategies:** Continued research on progressive loss strategies, as discussed in Chapter 6, can focus on refining these methods to achieve even better noise reduction and speech quality. Investigating different combinations of progressive loss functions and their application to various neural network architectures could yield valuable results.
- **Scalability and Real-Time Application:** Future research should address the scalability of the proposed methods for real-time applications. This includes optimizing computational efficiency to ensure that advanced speech enhancement techniques can be deployed effectively in real-world scenarios, such as mobile devices and embedded systems.
- **Multimodal Speech Enhancement:** Exploring the integration of multimodal data (e.g., combining audio with visual cues) to enhance speech signals. This could be particularly useful in scenarios where visual information is available, such as video calls or augmented reality environments.
- **User-Centric Adaptation:** Developing adaptive systems that can tailor speech enhancement parameters or models based on individual user preferences and specific use cases. This personalization can significantly improve user experience and satisfaction.

By pursuing these future research directions, we can continue to build on the foundational work presented in this thesis, driving further advancements in speech enhancement technology and its practical applications.

Chapter 8

Conclusiones

En esta tesis, nos propusimos mejorar el realce del habla utilizando redes neuronales profundas, centrándonos en el desarrollo y evaluación de nuevos métodos que superen las técnicas existentes. Las principales contribuciones han sido introducir la arquitectura de *Wide Residual Network* (WRN) para la mejora del habla, mejorar la interpretabilidad del modelo y explorar estrategias de coste progresivo.

Nuestra investigación demostró que la arquitectura WRN supera significativamente a los métodos tradicionales en diversos entornos ruidosos. Además, las técnicas de visualización proporcionaron valiosas percepciones sobre el comportamiento de las redes neuronales, lo que llevó al desarrollo de estrategias de fución de coste progresivo efectivas que mejoran la calidad del habla.

Los hallazgos contribuyen al campo al ofrecer una arquitectura robusta para la mejora del habla que equilibra el rendimiento y la complejidad computacional. Nuestros avances en interpretabilidad ayudan a cerrar la brecha entre los modelos de "caja negra" y la comprensión del usuario, proporcionando más confianza en las aplicaciones de redes neuronales.

8.1 Mejora del Habla Basada en WRN

En el Capítulo 4, la Mejora del Habla Basada en WRN introdujo un enfoque novedoso para la mejora del habla utilizando una arquitectura de *Wide Residual Network* (WRN). Al aprovechar las representaciones

proporcionadas por una topología *Wide* de *Convolutional Neural Networks* (CNNs) con conexiones residuales, este método demostró mejoras significativas sobre los métodos actuales más avanzados, particularmente en el manejo del habla con reverberación de campo lejano. Los siguientes puntos resumen los hallazgos y perspectivas clave:

- **Rendimiento Superior:** El método WRN superó significativamente al método *RNN-LSTM-based Weighted Prediction Error* (WPE), particularmente en el manejo del habla con reverberación de campo lejano en varias tamaños de habitaciones. Estos resultados prometedores y las mejoras metodológicas proporcionan una base sólida para futuras investigaciones, asegurando la relevancia continua y la eficacia del método WRN en la mejora del habla.
- **Beneficios de las Conexiones Residuales:** La inclusión de conexiones residuales permitió que la red mantuviera vías de señal lineales mientras mejoraba las correcciones no lineales, mejorando el rendimiento de la aplicación práctica.
- **Visualización para Mejorar la Interpretabilidad:** Las técnicas de visualización en capítulos posteriores ayudan a comprender el proceso de mejora del modelo WRN, abordando la naturaleza de "caja negra" del aprendizaje profundo e identificando pasos clave para la optimización.

En conclusión, el método WRN presentado en este capítulo ofrece una solución robusta y efectiva para la mejora del habla, superando a los métodos existentes en condiciones desafiantes. Los capítulos subsiguientes proporcionan análisis detallados y mejoras metodológicas que mejoran tanto el rendimiento como la interpretabilidad del modelo. Estos avances aseguran que el método WRN siga siendo una base sólida para investigaciones y desarrollos futuros en el campo del realce del habla, culminando en la publicación de estos hallazgos en:

- **J. Llombart**, D. Ribas, A. Miguel, L. Vicente, A. Ortega, y E. Lleida, **"Speech Enhancement with Wide Residual Networks in Reverberant Environments"** Proc. Interspeech 2019, 2019, pp. 1811–1815.

8.2 Técnicas de Visualización en la Mejora del Habla

En el Capítulo 5, la Visualización se centra en mejorar la interpretabilidad del realce del habla a través de arquitecturas de aprendizaje profundo. La exploración de nuevas arquitecturas apuntaba a hacer el proceso de mejora más transparente y comprensible. Al emplear diversas técnicas de visualización, el capítulo buscaba desmitificar la naturaleza de "caja negra" de las redes neuronales y mejorar el diseño general y la efectividad de las soluciones de realce del habla. A continuación, se resumen los principales hallazgos y percepciones:

- **Arquitecturas Propuestas:** El capítulo presentó nuevas arquitecturas, incluyendo la *Constant Channel Residual Network* (CCRN) y la *Constant Channel Residual Network with State Path* (CCRN-State), que mejoraron el rendimiento y la interpretabilidad de la mejora del habla. Estas arquitecturas mostraron resultados prometedores en la mejora de la calidad del habla y proporcionaron valiosas percepciones sobre el funcionamiento interno de los modelos, sugiriendo un buen equilibrio entre rendimiento e interpretabilidad.
- **Técnicas de Visualización:** Se emplearon sondas de visualización para monitorear y entender el proceso de mejora en cada bloque de la red. Esta supervisión paso a paso proporcionó percepciones sobre cómo diferentes componentes de la red contribuyen a la mejora general.
- **Supervisión Progresiva:** Se implementó el concepto de supervisión progresiva para seguir el proceso de mejora de manera incremental, lo que ayudó a identificar etapas críticas que impactan significativamente en la calidad del habla.

En conclusión, demostramos con éxito que integrar técnicas de visualización en arquitecturas de aprendizaje profundo para la mejora del habla puede mejorar significativamente la interpretabilidad y el rendimiento de los modelos. Las percepciones obtenidas de este estudio proporcionan una base sólida para futuras investigaciones y desarrollo en la creación de soluciones de mejora del habla más transparentes y efectivas. Los hallazgos de este capítulo han sido publicados en:

- **J. Llombart**, D. Ribas, A. Miguel, L. Vicente, A. Ortega, y E. Lleida, **"Progressive Speech Enhancement with Residual Connections"** Proc. Interspeech 2019, 2019, pp. 3193–3197.

8.3 Estrategias de Coste Progresivo en el Realce del Habla

En el Capítulo 6, exploramos estrategias de coste progresivo para el realce del habla utilizando arquitecturas de redes neuronales. El objetivo era mejorar el realce del habla mediante el ajuste del proceso de entrenamiento y la optimización del rendimiento a través de estrategias de coste innovadoras. Los siguientes puntos clave resumen las conclusiones obtenidas:

- **Estrategias de Coste Progresivo:** La implementación de estrategias de coste progresivo, incluyendo los métodos *Weighted Progressive* (WP) y *Uniform Progressive* (UP), mejoró significativamente la capacidad del modelo para realzar la calidad del habla al reducir sistemáticamente el ruido y la reverberación a través de la red.
- **Análisis de Gradientes:** Estudios preliminares de los gradientes indicaron que las estrategias de pérdida progresiva ayudan a estabilizar el proceso de entrenamiento en arquitecturas profundas, llevando a un rendimiento más consistente y robusto en diferentes condiciones de ruido y entornos.
- **Análisis de la Profundidad de la Arquitectura:** Las arquitecturas de redes neuronales más profundas, combinadas con estrategias de coste progresivo, mostraron un rendimiento mejorado en las medidas de calidad del habla, indicando la importancia de la profundidad para alcanzar resultados superiores.

En conclusión, esto destacó la efectividad de las estrategias de coste progresivo en la mejora de la calidad del habla. Estas estrategias proporcionan un marco robusto para avances futuros en la tecnología de realce del habla. Los hallazgos presentados subrayan el potencial de las estrategias de coste progresivo para establecer nuevos estándares

en el campo, aportando valiosas percepciones para futuras investigaciones e implementaciones prácticas. Los resultados de este estudio han sido publicados en:

- **J. Llombart**, D. Ribas, A. Miguel, L. Vicente, A. Ortega, y E. Lleida, **"Progressive loss functions for speech enhancement with deep neural networks"** EURASIP Journal on Audio, Speech, and Music Processing, vol. 2021, pp. 1–16, 2021.

8.4 Líneas Futuras de Investigación

Basándonos en los hallazgos y perspectivas prometedoras de esta tesis, se pueden perseguir varias direcciones de investigación futuras para avanzar aún más en el campo del realce del habla utilizando redes neuronales profundas. Los siguientes puntos esbozan áreas potenciales para futuras exploraciones y desarrollos:

- **Optimización de las Arquitecturas WRN:** Un refinamiento y optimización adicionales de las arquitecturas *Wide Residual Network* (WRN) podrían llevar a mejoras de rendimiento aún mayores. Esto incluye explorar diferentes configuraciones y hiperparámetros para mejorar su capacidad de manejar diversos entornos ruidosos de manera más efectiva.

- **Técnicas de Visualización Avanzadas:** Expandiendo los métodos de visualización introducidos en el Capítulo 5, trabajos futuros podrían desarrollar herramientas más sofisticadas para proporcionar percepciones más profundas sobre los procesos de mejora de las redes neuronales. Esto podría involucrar técnicas de visualización en tiempo real que permitan un monitoreo y ajuste dinámicos durante la fase de entrenamiento, por ejemplo, aprendiendo el número óptimo de pasos de limpieza de ruido en el marco progresivo.

- **Integración de Modelos Adicionales de Aprendizaje Profundo:** Incorporar otros modelos avanzados de aprendizaje profundo, como los *transformers* (Vaswani et al., 2017), modelos *state-space* (Gu & Dao, 2023), nuevos modelos recurrentes (Beck et al., 2024), frameworks de aprendizaje auto-supervisado (Chen et al., 2022)

y modelos generativos basados en *Diffusion* (Ho et al., 2020), podría proporcionar fortalezas complementarias a las arquitecturas WRN. Esta integración tiene el potencial de mejorar aún más la mejora del habla, especialmente en entornos acústicos más desafiantes, pero algunos métodos pueden generar habla nueva no presente en el audio fuente, por lo que prevenir esto es una de las principales líneas de investigación.

- **Exploración de Estrategias de Coste Progresivo:** La investigación continua sobre estrategias de coste progresivo, como se discutió en el Capítulo 6, puede centrarse en refinar estos métodos para lograr una mejor reducción de ruido y calidad del habla. Investigar diferentes combinaciones de funciones de coste progresivos y su aplicación a diversas arquitecturas de redes neuronales podría arrojar resultados valiosos.
- **Escalabilidad y Aplicación en Tiempo Real:** La investigación futura debería abordar la escalabilidad de los métodos propuestos para aplicaciones en tiempo real. Esto incluye optimizar la eficiencia computacional para garantizar que las técnicas avanzadas de mejora del habla se puedan implementar de manera efectiva en escenarios del mundo real, como dispositivos móviles y sistemas integrados.
- **Mejora del Habla Multimodal:** Explorar la integración de datos multimodales (por ejemplo, combinando señales de audio con señales visuales) para mejorar las señales del habla. Esto podría ser particularmente útil en escenarios donde la información visual está disponible, como en videollamadas o entornos de realidad aumentada.
- **Adaptación Centrada en el Usuario:** Desarrollar sistemas adaptables que puedan personalizar los parámetros o modelos de mejora del habla basados en las preferencias individuales del usuario y casos de uso específicos. Esta personalización puede mejorar significativamente la experiencia y satisfacción del usuario.

Al seguir estas direcciones de investigación futuras, podemos continuar construyendo sobre el trabajo fundacional presentado en esta tesis,

impulsando avances adicionales en la tecnología de realce del habla y sus aplicaciones prácticas.

Appendix A

STFT and Overlap-Add Method

The Short-Time Fourier Transform (STFT) and the overlap-add method are two fundamental techniques in speech enhancement that work together to process and reconstruct speech signals. The STFT transforms a time-domain signal into a time-frequency representation, allowing for detailed analysis and manipulation of the signal's spectral content. Once the speech enhancement process modifies the magnitude spectrum, the overlap-add method is employed to reconstruct the time-domain signal. By overlapping and adding the modified segments, this method ensures a smooth and continuous reconstruction, preserving the naturalness and intelligibility of the speech. Together, these techniques enable effective enhancement of speech signals by leveraging the frequency domain's detailed information and ensuring accurate time-domain reconstruction.

A.1 Short-Time Fourier Transform (STFT)

The Short-Time Fourier Transform (STFT) is a fundamental tool in signal processing for analyzing signals in the frequency domain. It transforms a time-domain signal $x(t)$ into a time-frequency representation by applying the Fourier transform to short, overlapping segments of the signal. Mathematically, the STFT is defined as:

$$X(t,f) = \int_{-\infty}^{\infty} x(\tau)w(\tau - t)e^{-j2\pi f\tau}d\tau \tag{A.1}$$

where $x(\tau)$ is the input signal, $w(\tau - t)$ is a window function centered at time t and $e^{-j2\pi f\tau}$ is the complex exponential function representing the Fourier transform. This allows for the examination of the signal's frequency content over time, which is crucial for analyzing non-stationary signals like speech (Oppenheim et al., 2011).

In speech enhancement, the STFT is used to decompose the speech signal into small, overlapping frames to capture transient features effectively. Each frame is windowed to reduce edge effects, and the Fourier transform is applied to convert it into the frequency domain. This process creates a spectrogram that represents the magnitude and phase of the signal's frequency components over time. Typically, the magnitude spectrum is used for enhancement because it contains most of the perceptual information, while the noisy phase is retained for reconstruction (Griffin & Lim, 1984).

In speech enhancement, only the magnitude of the STFT is typically used because it captures the essential features of the speech signal. The noisy phase is retained because accurate phase estimation is challenging, and imperfect phase reconstruction can introduce artifacts. Therefore, models enhance the magnitude spectrum while keeping the phase unchanged, ensuring a more natural reconstructed speech signal (Ephraim & Malah, 1984).

A.2 Overlap-Add

The overlap-add method is a technique used to reconstruct a time-domain signal from its Short-Time Fourier Transform (STFT) representation (Griffin & Lim, 1984). It involves overlapping and adding the inverse Fourier-transformed segments to synthesize the enhanced speech signal. This method ensures the continuity and naturalness of the reconstructed signal.

These are the two steps:

1. Inverse STFT: Compute the inverse STFT for each frame to convert it back from the frequency domain to the time domain. The inverse STFT for each frame $X(t, f)$ is given by:

$$x_n(t) = \int_{-\infty}^{\infty} X(t, f) e^{j2\pi ft} df \tag{A.2}$$

   where $x_n(t)$ is the time-domain signal for the n-th frame, and $e^{j2\pi ft}$ is the complex exponential function for the inverse Fouriere transform.

2. Overlap and Add: Overlap each time-domain segment according to the original segmentation, typically with a 50% overlap, and add the segments together to form the continuous signal. The procedure can be described as:

$$x(t) = \sum_n x_n(t - nH) \tag{A.3}$$

   where H is the hop size (i.e., the interval between the start of consecutive frames), and $x(t)$ is the reconstructed signal. The overlap helps to smooth the transitions between frames and reduce artifacts.

The overlap-add method is essential in reconstructing the enhanced speech signal from its STFT representation. By accurately overlapping and adding the inverse-transformed frames, it ensures that the enhanced speech maintains continuity and naturalness. This method effectively combines the enhanced magnitude spectrum with the retained noisy phase, resulting in a more intelligible and natural-sounding speech output. The overlap-add technique mitigates discontinuities and artifacts that might arise from the frame-based processing, preserving the integrity of the speech signal.

Appendix B

Results of noise experiment

This apendix show the complete sets of results obtained using the different architectures of the Chapter 6. This is also shown in Figure 6.6.

TABLE B.1: Results in Simulated REVERB-Eval set for different noises at different initial SNR.

SNR	Babble	Cafe	Music	Traffic	Tram	Average
			CNN (Estimated SNR / LLR)			
0	12.74 / 0.84	20.00 / 0.77	13.24 / 0.93	29.68 / 0.71	26.09 / 0.76	20.35 / 0.80
5	29.24 / 0.64	34.23 / 0.60	26.62 / 0.66	42.24 / 0.56	42.28 / 0.59	34.92 / 0.61
10	38.65 / 0.54	40.57 / 0.53	32.45 / 0.54	42.83 / 0.51	42.10 / 0.53	39.32 / 0.53
15	37.14 / 0.51	37.31 / 0.51	33.95 / 0.51	38.47 / 0.50	37.78 / 0.51	36.93 / 0.51
20	35.86 / 0.50	35.76 / 0.50	34.71 / 0.50	35.95 / 0.50	35.88 / 0.50	35.63 / 0.50
25	35.61 / 0.50	35.59 / 0.50	35.37 / 0.50	35.65 / 0.50	35.63 / 0.50	35.57 / 0.50
			P-CNN with WP (Estimated SNR / LLR)			
0	10.45 / 0.93	15.38 / 0.90	11.11 / 1.06	17.56 / 0.89	16.35 / 0.92	14.17 / 0.94
5	19.18 / 0.83	22.31 / 0.82	18.73 / 0.89	24.51 / 0.81	24.01 / 0.82	21.75 / 0.84
10	24.06 / 0.80	25.32 / 0.80	23.09 / 0.81	26.33 / 0.79	26.00 / 0.80	24.96 / 0.80
15	25.41 / 0.79	25.71 / 0.79	24.77 / 0.79	26.12 / 0.79	25.81 / 0.79	25.57 / 0.79
20	25.66 / 0.79	25.69 / 0.79	25.34 / 0.79	25.74 / 0.79	25.68 / 0.79	25.62 / 0.79
25	25.69 / 0.79	25.70 / 0.79	25.61 / 0.79	25.73 / 0.79	25.70 / 0.79	25.68 / 0.79
			P-CNN with UP (Estimated SNR / LLR)			
0	14.34 / 0.84	34.01 / 0.78	15.58 / 0.99	31.38 / 0.72	41.89 / 0.78	27.44 / 0.82
5	34.57 / 0.65	51.18 / 0.63	34.37 / 0.71	46.58 / 0.61	55.58 / 0.62	44.46 / 0.64
10	49.16 / 0.57	54.36 / 0.56	47.56 / 0.57	53.47 / 0.55	55.58 / 0.56	52.03 / 0.56
15	52.56 / 0.54	53.52 / 0.54	50.77 / 0.54	54.54 / 0.54	54.40 / 0.54	53.16 / 0.54
20	52.19 / 0.53	52.28 / 0.53	51.36 / 0.53	52.42 / 0.53	52.46 / 0.53	52.14 / 0.53
25	51.99 / 0.53	51.95 / 0.53	51.78 / 0.53	51.98 / 0.53	51.94 / 0.53	51.93 / 0.53
			ResNet (Estimated SNR / LLR)			
0	12.94 / 0.79	27.71 / 0.73	15.39 / 0.84	27.58 / 0.67	32.64 / 0.73	23.25 / 0.75
5	26.44 / 0.59	38.18 / 0.55	27.15 / 0.59	35.72 / 0.53	43.03 / 0.55	34.11 / 0.56
10	34.51 / 0.51	37.63 / 0.50	31.89 / 0.50	36.97 / 0.49	39.25 / 0.49	36.05 / 0.50
15	34.96 / 0.48	35.46 / 0.48	33.55 / 0.48	36.04 / 0.47	35.94 / 0.48	35.19 / 0.48
20	34.41 / 0.47	34.45 / 0.47	33.86 / 0.47	34.50 / 0.47	34.51 / 0.47	34.35 / 0.47
25	34.21 / 0.47	34.24 / 0.47	34.07 / 0.47	34.26 / 0.47	34.23 / 0.47	34.20 / 0.47
			P-ResNet with WP (Estimated SNR / LLR)			
0	15.22 / 0.86	32.58 / 0.79	14.14 / 1.02	29.04 / 0.75	39.48 / 0.79	26.09 / 0.84
5	38.97 / 0.64	57.25 / 0.60	36.74 / 0.70	51.67 / 0.59	62.66 / 0.61	49.46 / 0.63
10	53.23 / 0.52	57.93 / 0.51	49.75 / 0.53	58.03 / 0.50	59.30 / 0.51	55.65 / 0.51
15	54.03 / 0.49	55.02 / 0.48	51.38 / 0.49	56.45 / 0.48	55.36 / 0.48	54.45 / 0.48
20	53.71 / 0.48	53.73 / 0.48	52.31 / 0.48	54.07 / 0.48	53.80 / 0.48	53.52 / 0.48
25	53.60 / 0.48	53.58 / 0.48	53.28 / 0.48	53.68 / 0.48	53.56 / 0.48	53.54 / 0.48
			P-ResNet with UP (Estimated SNR / LLR)			
0	10.65 / 0.86	20.78 / 0.82	11.41 / 0.99	20.74 / 0.77	22.91 / 0.83	17.30 / 0.85
5	24.49 / 0.65	34.68 / 0.62	23.54 / 0.69	31.38 / 0.59	38.09 / 0.62	30.44 / 0.63
10	38.09 / 0.54	40.87 / 0.53	34.71 / 0.54	40.11 / 0.51	42.19 / 0.52	39.19 / 0.53
15	41.02 / 0.50	41.27 / 0.50	38.70 / 0.50	42.06 / 0.49	41.65 / 0.50	40.94 / 0.50
20	40.53 / 0.49	40.41 / 0.49	39.40 / 0.49	40.56 / 0.49	40.50 / 0.49	40.28 / 0.49
25	40.26 / 0.49	40.25 / 0.49	40.02 / 0.49	40.31 / 0.49	40.28 / 0.49	40.22 / 0.49

Bibliography

Loizou, P. C. (2007). *Speech enhancement: Theory and practice*. CRC press.

Nakatani, T., Yoshioka, T., Kinoshita, K., Miyoshi, M., & Juang, B.-H. (2010). Speech dereverberation based on variance-normalized delayed linear prediction, IEEE.

Hoffmann, R. (2010). On the development of early vocoders, In *2010 second region 8 ieee conference on the history of communications*. IEEE.

Boll, S. (1979a). Suppression of acoustic noise in speech using spectral subtraction. *IEEE Trans. on Acoustic, Speech and Signal Processing*, *27*(2), 113–120.

Makhoul, J. (1975). Linear prediction: A tutorial review. *Proceedings of the IEEE*, *63*(4), 561–580.

Atal, B. S. (1974). Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. *the Journal of the Acoustical Society of America*, *55*(6), 1304–1312.

Markel, J. D., & Gray, A. J. (1976). *Linear prediction of speech* (Vol. 12). Springer Communication; Cybernetics.

Paliwal, K., & Basu, A. (1987). A speech enhancement method based on kalman filtering, In *Icassp '87. ieee international conference on acoustics, speech, and signal processing*. https://doi.org/10.1109/ICASSP.1987.1169756

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems.

Brown, R. G. (1983). Introduction to random signal analysis and kalman filtering(book). *New York, John Wiley and Sons, 1983, 357 p.*

Boll, S. (1979b). Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on acoustics, speech, and signal processing*, *27*(2), 113–120.

Lim, J. S., & Oppenheim, A. V. (1979). Enhancement and bandwidth compression of noisy speech. *Proceedings of the IEEE*, *67*(12), 1586–1604.

Malah, D., Cox, R. V., & Accardi, A. J. (1999). Tracking speech-presence uncertainty to improve speech enhancement in non-stationary noise environments, In *1999 ieee international conference on acoustics, speech, and signal processing. proceedings. icassp99 (cat. no. 99ch36258)*. IEEE.

Cohen, I. (2003). Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging. *IEEE Transactions on speech and audio processing*, *11*(5), 466–475.

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

Ephraim, Y., & Malah, D. (1985). Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. *IEEE transactions on acoustics, speech, and signal processing*, *33*(2), 443–445.

Rabiner, L. R. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, *77*(2), 257–286.

Veisi, H., & Sameti, H. (2013). Speech enhancement using hidden markov models in mel-frequency domain. *Speech Communication*, *55*(2), 205–220.

Daubechies, I. (1990). The wavelet transform, time-frequency localization and signal analysis. *IEEE transactions on information theory*, *36*(5), 961–1005.

Mohammadiha, N., Smaragdis, P., & Leijon, A. (2013). Supervised and unsupervised speech enhancement using nonnegative matrix factorization. *IEEE Transactions on audio, speech, and language processing*, *21*(10), 2140–2151.

Fan, H.-T., Hung, J.-w., Lu, X., Wang, S.-S., & Tsao, Y. (2014). Speech enhancement using segmental nonnegative matrix factorization, In *2014 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE.

Wan, E. A., Nelson, A. T., Katagiri, S., Et al. (1999). Networks for speech enhancement. *Handbook of neural networks for speech processing. Artech House, Boston, USA*, *139*(1), 7.

Maas, A. L., Le, Q. V., O'Neil, T. M., Vinyals, O., Nguyen, P., & Ng, A. Y. (2012). Recurrent neural networks for noise reduction in robust asr., In *Interspeech*.

Pascual, S., Bonafonte, A., & Serra, J. (2017). Segan: Speech enhancement generative adversarial network. *arXiv preprint arXiv:1703.09452*.

Mowlaee, P., & Kulmer, J. (2015). Phase estimation in single-channel speech enhancement: Limits-potential. *IEEE Transactions Audio, Speech, and Language Processing*, *23*(8), 1283–1294.

Mowlaee, P., Kulmer, J., Stahl, J., & Mayer, F. (2016). *Single channel phase-aware signal processing in speech communication: Theory and practice*. Wiley.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Oostermeijer, K., Wang, Q., & Du, J. (2021). Lightweight causal transformer with local self-attention for real-time speech enhancement., In *Interspeech*.

Qiu, Y., Wang, R., Singh, S., Ma, Z., & Hou, F. (2021). Self-supervised learning based phone-fortified speech enhancement., In *Interspeech*.

Huang, Z., Watanabe, S., Yang, S.-w., García, P., & Khudanpur, S. (2022). Investigating self-supervised learning for speech enhancement and separation, In *Icassp 2022-2022 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE.

Wang, D. (2005). On ideal binary mask as the computational goal of auditory scene analysis, In *Speech separation by humans and machines*. Springer.

Williamson, D. S., Wang, Y., & Wang, D. (2015). Complex ratio masking for monaural speech separation. *IEEE/ACM transactions on audio, speech, and language processing*, *24*(3), 483–492.

Miller, G. A. (1947). The masking of speech. *Psychological bulletin*, *44*(2), 105.

Yilmaz, O., & Rickard, S. (2004). Blind separation of speech mixtures via time-frequency masking. *IEEE Transactions on signal processing*, *52*(7), 1830–1847.

Narayanan, A., & Wang, D. (2013). Ideal ratio mask estimation using deep neural networks for robust speech recognition, In *2013 ieee international conference on acoustics, speech and signal processing*. IEEE.

Ribas, D., Miguel, A., Ortega, A., & Lleida, E. (2022). Wiener filter and deep neural networks: A well-balanced pair for speech enhancement. *Applied Sciences*, *12*(18), 9000.

Berouti, M., Schwartz, R., & Makhoul, J. (1979). Enhancement of speech corrupted by acoustic noise, In *Icassp'79. ieee international conference on acoustics, speech, and signal processing*. IEEE.

Griffin, D., & Lim, J. (1984). Signal estimation from modified short-time fourier transform. *IEEE Transactions on acoustics, speech, and signal processing*, *32*(2), 236–243.

Xu, Y., Du, J., Dai, L. R., & Lee, C. H. (2014). An experimental study on speech enhancement based on deep neural networks. *IEEE Signal Processing Letters*, *21*(1), 65–68.

Bando, Y., Mimura, M., Itoyama, K., Yoshii, K., & Kawahara, T. (2018). Statistical speech enhancement based on probabilistic integration of variational autoencoder and non-negative matrix factorization, In *2018 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE.

Qian, K., Zhang, Y., Chang, S., Yang, X., Florencio, D., & Hasegawa-Johnson, M. (2017). Speech enhancement using bayesian wavenet, In *Interspeech*.

Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., Manzagol, P.-A., & Bottou, L. (2010). Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research*, *11*(12).

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems*, *27*.

Van Den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., Kavukcuoglu, K., Et al. (2016). Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, *12*.

Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in neural information processing systems*, *33*, 6840–6851.

Gonzalez, P., Tan, Z.-H., Østergaard, J., Jensen, J., Alstrøm, T. S., & May, T. (2024). Diffusion-based speech enhancement in matched and mismatched conditions using a heun-based sampler, In *Icassp 2024-2024 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE.

Snyder, D., Chen, G., & Povey, D. (2015). MUSAN: A Music, Speech, and Noise Corpus [arXiv:1510.08484v1].

Dean, J., Corrado, G., Monga, R., Chen, K., Devin, M., Mao, M., Ranzato, M., Senior, A., Tucker, P., Yang, K., Et al. (2012). Large scale distributed deep networks. *Advances in neural information processing systems*, *25*.

Diaz-Guerra, D., Miguel, A., & Beltran, J. R. (2021). Gpurir: A python library for room impulse response simulation with gpu acceleration. *Multimedia Tools and Applications*, *80*(4), 5653–5671.

Ko, T., Peddinti, V., Povey, D., & Khudanpur, S. (2015). Audio augmentation for speech recognition., In *Interspeech*.

Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., Prenger, R., Satheesh, S., Sengupta, S., Coates, A., Et al. (2014). Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*.

Virtanen, T., Singh, R., & Raj, B. (2012). *Techniques for noise robustness in automatic speech recognition*. John Wiley & Sons.

Naylor, P. A., & Gaubitch, N. D. (2010). *Speech dereverberation*. Springer Science & Business Media.

Allen, J. B., & Berkley, D. A. (1979a). Image method for efficiently simulating small-room acoustics. *The Journal of the Acoustical Society of America*, *65*(4), 943–950.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press.

Bishop, C. M. (2006). Pattern recognition and machine learning. *Springer google schola*, *2*, 1122–1128.

Huber, P. J. (1992). Robust estimation of a location parameter, In *Breakthroughs in statistics: Methodology and distribution*. Springer.

Siniscalchi, S. M. (2021). Vector-to-vector regression via distributional loss for speech enhancement. *IEEE Signal Processing Letters*, *28*, 254–258.

Mermelstein, P. (1976). Distance measures for speech recognition, psychological and instrumental. *Pattern recognition and artificial intelligence*, *116*, 374–388.

Davis, S., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE transactions on acoustics, speech, and signal processing*, *28*(4), 357–366.

Hermansky, H. (1990). Perceptual linear predictive (plp) analysis of speech. *the Journal of the Acoustical Society of America*, *87*(4), 1738–1752.

Hermansky, H., & Morgan, N. (1994). Rasta processing of speech. *IEEE transactions on speech and audio processing*, *2*(4), 578–589.

Hu, Y., & Loizou, P. C. (2007). Evaluation of objective quality measures for speech enhancement. *IEEE Transactions on audio, speech, and language processing*, *16*(1), 229–238.

Gray, A., & Markel, J. (1976). Distance measures for speech processing. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *24*(5), 380–391.

Falk, T. H., Zheng, C., & Chan, W.-Y. (2010b). A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech. *IEEE Transactions on Audio, Speech, and Language Processing*, *18*(7), 1766–1774.

Rix, A. W., Beerends, J. G., Hollier, M. P., & Hekstra, A. P. (2001). Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs, In *2001 ieee international conference on acoustics, speech, and signal processing. proceedings (cat. no. 01ch37221)*. IEEE.

ITU-T Recommendation. (2001). Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs. *Rec. ITU-T P. 862*.

Taal, C. H., Hendriks, R. C., Heusdens, R., & Jensen, J. (2011). An algorithm for intelligibility prediction of time–frequency weighted noisy speech. *IEEE Transactions on audio, speech, and language processing*, *19*(7), 2125–2136.

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition, In *The ieee conference on computer vision and pattern recognition (cvpr)*.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735–1780.

Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *science*, *313*(5786), 504–507.

Deng, L., & Li, X. (2013). Machine learning paradigms for speech recognition: An overview. *IEEE Transactions on Audio, Speech, and Language Processing*, *21*(5), 1060–1089.

Fu, S.-W., Tsao, Y., & Lu, X. (2016). SNR-aware convolutional neural network modeling for speech enhancement, In *Interspeech*.

Park, S. R., & Lee, J. (2017). A fully convolutional neural network for speech enhancement, In *Interspeech*.

Weninger, F., Erdogan, H., Watanabe, S., Vincent, E., Roux, J. L., Hershey, J. R., & Schuller, B. (2015). Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR, In *Proceedings of the international conference on latent variable analysis and signal separation*.

Chen, J., & Wang, D. (2016). Long short-term memory for speaker generalization in supervised speech separation, In *Interspeech*.

Kinoshita, K., Delcroix, M., Kwon, H., Mori, T., & Nakatani, T. (2017). Neural network-based spectrum estimation for online WPE dereverberation, In *Interspeech*.

Gao, T., Du, J., Dai, L.-R., & Lee, C.-H. (2018). Densely connected progressive learning for LSTM-based speech enhancement, In *Ieee international conference on acoustic, speech and signal processing (icassp)*.

Santos, J. F., & Falk, T. H. (2018). Speech dereverberation with context-aware recurrent neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *26*(7), 1236–1246.

Zagoruyko, S., & Komodakis, N. (2017). Wide residual networks. *CoRR*, *abs/1605.07146*arXiv. https://arxiv.org/abs/1605.07146

Llombart, J., Miguel, A., Ortega, A., & Lleida, E. (2018). Wide residual networks 1d for automatic text punctuation. *IberSPEECH 2018*, 296–300.

Kinoshita, K., Delcroix, M., Yoshioka, T., Nakatani, T., Habets, E., Haeb-Umbach, R., Leutnant, V., Sehr, A., Kellermann, W., Maas, R.,

Gannot, S., & Raj, B. (2013). The REVERB Challenge: A common evaluation framework for dereverberation and recognition of reverberant speech, In *Proceedings of the ieee workshop on applications of signal processing to audio and acoustics (waspaa-13)*.

Robinson, T., Fransen, J., Pye, D., Foote, J., & Renals, S. (1995). WSJ-CAMO: a British English speech corpus for large vocabulary continuous speech recognition, In *Ieee international conference on acoustic, speech and signal processing (icassp)*.

Lincoln, M., McCowan, I., Vepa, J., & Maganti, H. (2005). The multichannel Wall Street Journal audio visual corpus (MC-WSJ-AV): specification and initial experiments, In *Proceedings of the 2005 ieee workshop on automatic speech recognition and understanding (asru-05)*.

Bertin, N., Camberlein, E., Vincent, E., Lebarbenchon, R., Peillon, S., Éric Lamandé, Sivasankaran, S., Bimbot, F., Illina, I., Tom, A., Fleury, S., & Éric Jamet. (2016). A french corpus for distant-microphone speech processing in real homes, In *Interspeech*.

Bertin, N., Camberlein, E., Lebarbenchon, R., Vincent, E., Sivasankaran, S., Illina, I., & Bimbot, F. (2019). VoiceHome-2, an extended corpus for multichannel speech processing in real homes. *Speech Communications*, *106*, 68 –78.

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., & Pallett, D. S. (1993). Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1. *NASA STI/Recon technical report n*, *93*.

Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015). Librispeech: an ASR corpus based on public domain audio books, In *2015 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE.

Rousseau, A., Deléglise, P., & Esteve, Y. (2014). Enhancing the TED-LIUM Corpus with Selected Data for Language Modeling and More TED Talks., In *Lrec*.

Loizou, P. C. (2011). Speech quality asssessment. in: Multimedia analysis, processing and communications. Springer.

Ramirez, J., Segura, J. C., Benitez, C., de la Torre, A., & Rubio, A. (2004). Efficient voice activity detection algorithms using long-term speech information. *Speech Communication*, *42*(3), 271 –287.

Falk, T. H., Zheng, C., & Chan, W.-Y. (2010a). A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech. *IEEE Transaction in Audio, Speech and Language Processing*, *18*(7), 1766–1774.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, In *Proceedings of the ieee international conference on computer vision*.

Drude, L., Heymann, J., Boeddeker, C., & Haeb-Umbach, R. (2018). NARA-WPE: A Python package for weighted prediction error dereverberation in Numpy and Tensorflow for online and offline processing, In *13. itg fachtagung sprachkommunikation (itg 2018)*.

Ribas, D., Vincent, E., & Calvo, J. R. (2016). A study of speech distortion conditions in real scenarios for speech processing applications, In *Ieee spoken language technology workshop (slt)*.

Xia, B.-Y., & Bao, C.-C. (2013). Speech enhancement with weighted denoising auto-encoder, In *Interspeech*.

Feng, X., Zhang, Y., & Glass, J. (2014). Speech feature denoising and dereverberation via deep autoencoders for noisy reverberant speech recognition, In *Ieee international conference on acoustic, speech and signal processing (icassp)*.

Tu, M., & Zhang, X. (2017). Speech enhancement based on deep neural networks with skip connections, In *Ieee international conference on acoustic, speech and signal processing (icassp)*.

Karjol, P., Kumar, A., & Ghosh, P. K. (2018). Speech enhancement using multiple deep neural networks, In *Ieee international conference on acoustic, speech and signal processing (icassp)*.

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *CoRR*, *abs/1502.03167*arXiv 1502.03167. http://arxiv.org/abs/1502.03167

Zagoruyko, S., & Komodakis, N. (2016). Wide residual networks. *CoRR*, *abs/1605.07146*arXiv 1605.07146. http://arxiv.org/abs/1605.07146

Lee, C.-Y., Xie, S., Gallagher, P., Zhang, Z., & Tu, Z. (2015). Deeply-supervised nets, In *Artificial intelligence and statistics*. Pmlr.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system, In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*.

Kingma, D. P., & Ba, J. L. (2015). Adam: Amethod for stochastic optimization, In *Proceedings of the 3rd international conference on learning representations (iclr)*.

Loshchilov, I., & Hutter, F. (2017). Fixing weight decay regularization in adam. *arXiv preprint arXiv:1711.05101*.

Gao, T., Du, J., Dai, L.-R., & Lee, C.-H. (2016). SNR-Based Progressive Learning of Deep Neural Network for Speech Enhancement., In *Interspeech*.

Llombart, J., Ribas, D., Miguel, A., Vicente, L., Ortega, A., & Lleida, E. (2019a). Progressive Speech Enhancement with Residual Connections, In *Proc. interspeech 2019*. https://doi.org/10.21437/Interspeech.2019-1748

Zhao, H., Zarar, S., Tashev, I., & Lee, C.-H. (2018). Convolutional-recurrent neural networks for speech enhancement, In *Ieee international conference on acoustic, speech and signal processing (icassp)*.

Chen, Z., Huang, Y., Li, J., & Gong, Y. (2017). Improving mask learning based speech enhancement system with restoration layers and residual connection, In *Interspeech*.

Llombart, J., Ribas, D., Miguel, A., Vicente, L., Ortega, A., & Lleida, E. (2019b). Speech Enhancement with Wide Residual Networks in Reverberant Environments, In *Proc. interspeech 2019*. https://doi.org/10.21437/Interspeech.2019-1745

Allen, J. B., & Berkley, D. A. (1979b). Image method for efficiently simulating small-room acoustics. *The Journal of the Acoustical Society of America*, *65*(4), 943–950.

Kim, C., & Stern, R. M. (2008). Robust signal-to-noise ratio estimation based on waveform amplitude distribution analysis, In *Ninth annual conference of the international speech communication association*.

Santos, J. F., Senoussaoui, M., & Falk, T. H. (2014). An updated objective intelligibility estimation metric for normal hearing listeners under noise and reverberation, In *Proc. int. workshop acoust. signal enhancement*.

Gelderblom, F. B., Tronstad, T. V., & Viggen, E. M. (2018). Subjective evaluation of a noise-reduced training target for deep neural

network-based speech enhancement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *27*(3), 583–594.

Santos, J. F., & Falk, T. H. (2019). Towards the development of a non-intrusive objective quality measure for dnn-enhanced speech, In *2019 eleventh international conference on quality of multimedia experience (qomex)*. IEEE.

Gu, A., & Dao, T. (2023). Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*.

Beck, M., Pöppel, K., Spanring, M., Auer, A., Prudnikova, O., Kopp, M., Klambauer, G., Brandstetter, J., & Hochreiter, S. (2024). Xlstm: Extended long short-term memory. *arXiv preprint arXiv:2405.04517*.

Chen, S., Wang, C., Chen, Z., Wu, Y., Liu, S., Chen, Z., Li, J., Kanda, N., Yoshioka, T., Xiao, X., Et al. (2022). Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, *16*(6), 1505–1518.

Oppenheim, A. V., Schafer, R. W., Buck, J. R., Et al. (2011). Tratamiento de señales en tiempo discreto.